

# Folding of Protein L with implications for collapse in the denatured state ensemble

Hiranmay Maity and Govardhan Reddy\*

*Solid State and Structural Chemistry Unit,*

*Indian Institute of Science, Bangalore, Karnataka, India 560012*

## Abstract

A fundamental question in protein folding is whether the coil to globule collapse transition occurs during the initial stages of folding (burst-phase) or simultaneously with the protein folding transition. Single molecule fluorescence resonance energy transfer (FRET) and small angle X-ray scattering (SAXS) experiments disagree on whether Protein L collapse transition occurs during the burst-phase of folding. We study Protein L folding using a coarse-grained model and molecular dynamics simulations. The collapse transition in Protein L is found to be concomitant with the folding transition. In the burst-phase of folding, we find that FRET experiments overestimate radius of gyration,  $R_g$ , of the protein due to the application of Gaussian polymer chain end-to-end distribution to extract  $R_g$  from the FRET efficiency. FRET experiments estimate  $\approx 6\text{\AA}$  decrease in  $R_g$  when the actual decrease is  $\approx 3\text{\AA}$  on Guanidinium Chloride denaturant dilution from 7.5M to 1M, and thereby suggesting pronounced compaction in the protein dimensions in the burst-phase. The  $\approx 3\text{\AA}$  decrease is close to the statistical uncertainties of the  $R_g$  data measured from SAXS experiments, which suggest no compaction, leading to a disagreement with the FRET experiments. The transition state ensemble (TSE) structures in Protein L folding are globular and extensive in agreement with the  $\Psi$ -analysis experiments. The results support the hypothesis that the TSE of single domain proteins depend on protein topology, and are not stabilised by local interactions alone.

# Introduction

There is an ongoing debate<sup>1-5</sup> on whether the denatured ensemble of single domain proteins undergoes a coil to globule transition during the burst-phase of folding as the denaturant concentration is diluted to lower values. Proteins are heteropolymers and behave like random coils at high temperatures or denaturant concentrations<sup>6-8</sup>. An interesting question is whether proteins akin to polymers undergo a collapse transition in the burst-phase of folding as the conditions are made conducive for folding<sup>9-11</sup>. Single domain proteins unlike polymers are finite sized, and are composed of a specific sequence of amino acids which are hydrophobic and hydrophilic in character. The finite size effects and heteropolymer character are the reasons attributed to the marginal stability of proteins, and the near overlap of the collapse and folding transition temperatures, which makes them fold efficiently<sup>12,13</sup>.

The collapse transition in proteins is generally studied using single molecule fluorescence resonance energy transfer (FRET), and small angle X-ray scattering (SAXS) experiments. Although FRET and SAXS experiments agree that proteins like Cytochrome *c*<sup>14-17</sup> and Monellin<sup>18-20</sup> collapse during the burst-phase of folding, their results disagree for Protein L. FRET experiments for Protein L<sup>21-23</sup> infer collapse, whereas SAXS experiments<sup>24,25</sup> conclude no collapse in the burst-phase of folding on dilution of Guanidine Hydrochloride [*GuHCl*]. Both FRET and SAXS estimate the radius of gyration,  $R_g$ , of the protein to infer the size of the protein in the unfolded ensemble. The difference in the  $R_g$  predictions of FRET<sup>21,22</sup> and SAXS<sup>24,25</sup> experiments for Protein L during the burst-phase of folding is statistically significant. The reasons for the disagreement between these experiments for Protein L are not completely clear. Understanding the impact of various approximations used in these methods to estimate the size of the protein can not only aid in resolving the disagreement between FRET and SAXS but also to understand the problem of protein collapse better.

Single domain proteins close to the melting temperature or the mid-point denaturant concentration generally fold in a 2-state manner through an ensemble of transition state structures (TSE).  $\phi$ -analysis experiments for Protein L<sup>26-29</sup> predict that the TSE is polarised with only the N-terminal  $\beta$ -hairpin present. Whereas the  $\psi$ -analysis experiments<sup>30</sup> predict that the TSE is globular and extensive with both the N and C-termini  $\beta$ -hairpins present along with some non-native interactions in the C-terminal  $\beta$ -hairpin. The  $\psi$ -analysis experiments support that the TSE and the folding pathways for the single domain proteins

depend on the protein topology<sup>31</sup>, whereas the  $\phi$ -analysis experiments conclude that the TSE is mostly stabilised by local interactions<sup>32</sup>.

Various aspects of Protein L folding such as folding pathways, transition state structures, and properties of the unfolded ensemble are studied<sup>33–40</sup> using both coarse-grained and atomistic simulations. In this manuscript, we study the burst-phase folding of Protein L to understand the origin of discrepancy between the FRET<sup>21,22</sup> and SAXS<sup>24,25</sup> experiments using the native-centric self-organised polymer model with side chains (SOP-SC)<sup>41,42</sup> and molecular dynamics simulations. The effect of  $[GuHCl]$  on Protein L conformations is taken into account using the molecular transfer model (MTM)<sup>11,43</sup>.

The computed FRET efficiency,  $\langle E \rangle$ , for Protein L in the burst-phase of folding is in quantitative agreement with the FRET experiments of Eaton *et al.*<sup>22</sup>, and only in partial agreement with the experiments of Haran *et al.*<sup>21</sup>. The FRET experiments are found to overestimate  $R_g$  compared to the actual values computed directly from the simulations owing to the use of the Gaussian polymer chain end-to-end distribution function to extract  $R_g$  from  $\langle E \rangle$ . The deviation between FRET-extracted and actual  $R_g$  increased with  $[GuHCl]$ . As a result, FRET experiments<sup>22</sup> estimate  $\approx 6\text{\AA}$  decrease in  $R_g$  and infer protein collapse in the burst-phase, when the actual decrease is  $\approx 3\text{\AA}$  as  $[GuHCl]$  is diluted from 7.5M to 1M.

The equilibrium  $R_g$  computed as a function of  $[GuHCl]$  is in near quantitative agreement with the SAXS experiments<sup>25</sup>. The SAXS experiments<sup>25</sup> infer no protein collapse as the burst-phase  $R_g$  at  $[GuHCl] = 4.0\text{M}$  and  $0.67\text{M}$  are not statistically different. In the simulations, the burst-phase  $R_g$  at  $[GuHCl] = 4\text{M}$  and  $1\text{M}$  are  $25.1 \pm 4.0\text{\AA}$  and  $23.8 \pm 3.9\text{\AA}$ , respectively, a difference of  $\approx 1.3\text{\AA}$ , which is well within the standard deviation  $\sigma_{R_g} \approx 4\text{\AA}$ . From this analysis, which is similar to the SAXS analysis, we can infer no collapse as the change in protein dimensions are not statistically significant, leading to a disagreement with the FRET experiments.

The TSE of Protein L at the melting temperature is inferred using  $P_{fold}$  calculations<sup>44</sup>. The TSE is found to be globular and extensive with both the N and C termini  $\beta$ -hairpins present resembling a topology similar to that of the folded structure. The results are in agreement with the  $\Psi$ -analysis experiments<sup>30</sup> and only in partial agreement with the  $\phi$ -analysis experiments<sup>26–29</sup>. The inferred TSE support the hypothesis that the transition state structures of single domain homologous proteins are extensive and depend on the protein topology<sup>45</sup>.

# Methods

**Self Organised polymer-Side Chain (SOP-SC) Model:** We used the SOP-SC (self-organized polymer-side chain) model<sup>41,42</sup> in which each amino acid residue is represented by two beads. One bead is at the  $C_\alpha$  position representing the backbone atoms, and the other bead is at the center of mass of the side chain representing the side chain atoms. The effective energy of a protein conformation in the SOP-SC model is a sum of bonded and non-bonded interactions. The bonded interactions ( $E_B$ ) are present between a pair of connected beads. The non-bonded interactions are a sum of native ( $E_{NB}^N$ ) and non-native ( $E_{NB}^{NN}$ ) interactions (see Supporting Information (SI) for more details). The native interactions for protein L are identified using the crystal structure<sup>46</sup> (Protein Data Bank ID: 1HZ6) (Fig 1A). The number of residues in the crystal structure,  $N_{res} = 64$ . The native interactions between the beads representing the amino acid side chains interact via a residue dependent Betancourt-Thirumalai statistical potential<sup>47</sup>.

The coarse-grained force-field in the SOP-SC model for a protein conformation given by the co-ordinates  $\{\mathbf{r}\}$  in the absence of denaturants,  $[C] = 0$ , is

$$E_{CG}(\{\mathbf{r}\}, 0) = E_B + E_{NB}^N + E_{NB}^{NN}. \quad (1)$$

Description of the various energy terms in Equation 1 and the parameters used in the energy function are given in the SI. These parameters are identical to the values previously used to successfully study the folding properties of the proteins Ubiquitin<sup>48</sup> and GFP<sup>49</sup>. We used the same force-field to study the properties of different proteins, and as a result this force-field satisfies the criterion of a transferable force-field.

**Molecular Transfer Model:** To simulate Protein L folding thermodynamics and kinetics in the presence of  $[GuHCl]$  we used the Molecular Transfer Model (MTM)<sup>11,43</sup>. In the presence of a denaturant of concentration  $[C]$ , the effective coarse-grained force field for the protein using MTM is given by

$$E_{CG}(\{\mathbf{r}\}, [C]) = E_{CG}(\{\mathbf{r}\}, 0) + \Delta G_{tr}(\{\mathbf{r}\}, [C]), \quad (2)$$

where  $E_{CG}(\{\mathbf{r}\}, 0)$  is given by Eq. 1,  $\Delta G_{tr}(\{\mathbf{r}\}, [C])$  is the protein-denaturant interaction

energy in a solution with denaturant concentration  $[C]$ , and is given by

$$\Delta G_{tr}(\{\mathbf{r}\}, [C]) = \sum_{k=1}^N \delta g_{tr,k}([C]) \alpha_k(\{\mathbf{r}\}) / \alpha_{Gly-k-Gly}, \quad (3)$$

where  $N(=N_{res} \times 2 = 128)$  is the number of beads in coarse-grained Protein L,  $\delta g_{tr,k}([C])$  is the transfer free energy of bead  $k$ ,  $\alpha_k(\{\mathbf{r}\})$  is the solvent accessible surface area (SASA) of the bead  $k$  in a protein conformation described by positions  $\{\mathbf{r}\}$ ,  $\alpha_{Gly-k-Gly}$  is the SASA of the bead  $k$  in the tripeptide  $Gly - k - Gly$ . The radii for amino acid side chains to compute  $\alpha_k(\{\mathbf{r}\})$  are given in Table S2 in Ref.<sup>48</sup>. The experimental<sup>11,50,51</sup> transfer free energies  $\delta g_{tr,i}([C])$ , which depend on the chemical nature of the denaturant, for backbone and side chains are listed in Table S3 in Ref.<sup>42</sup>. The values for  $\alpha_{Gly-k-Gly}$  are listed in Table S4 in Ref.<sup>42</sup>.

**Simulations and Data Analysis:** Low friction Langevin dynamics simulations<sup>52</sup> are used to generate protein conformations as a function of  $T$  in  $[C] = 0M$  conditions. To compute thermodynamic properties of the protein in a denaturant solution of concentration  $[C]$ ,  $\Delta G_{tr}(\{\mathbf{r}\}, [C])$  is treated as perturbation to  $E_{CG}(\{\mathbf{r}\}, 0)$  in Eq. 2, and Weighted Histogram Method<sup>11,43,53</sup> is used to compute average value of various physical quantities at any  $[C]$ . Brownian dynamics simulations<sup>54</sup> are used with the full Hamiltonian (Eq. 2) to simulate the burst-phase folding kinetics of the protein in a denaturant solution of concentration  $[C]$  (see SI for details).

We computed structural overlap function<sup>55</sup>,  $\chi$ , and radius of gyration,  $R_g$ , to monitor protein L folding kinetics. The structural overlap function is defined as  $\chi = 1 - \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} \Theta(\delta - |r_i - r_i^0|)$ . Here,  $N_{tot}(= 777)$  is the number of pairs of beads in the SOP-SC model of Protein L assuming that the bead centers are separated by at least 2 bonds,  $r_i$  is the distance between the  $i^{th}$  pair of beads, and  $r_i^0$  being the corresponding distance in the folded state,  $\Theta$  is the Heaviside step function, and  $\delta = 2\text{\AA}$ . Using  $\chi$  as an order parameter, we calculated the fraction of molecules in the NBA,  $f_{NBA}$  as a function of  $[GuHCl]$  (see SI for details and Fig. S2).  $R_g$  is calculated using  $R_g = (1/2N^2)(\sum_{i,j} \vec{r}_{ij}^2)^{1/2}$ , where  $\vec{r}_{ij}$  is the vector connecting the beads  $i$  and  $j$ . The extent of long-range contacts in the TSE structures compared to the coarse-grained PDB structure is analysed using the relative contact order<sup>56,57</sup>,  $RCO$ , which is defined as  $RCO = \frac{1}{N_{res}N_{nat}} \sum_{i=1}^{N_{nat}} L_i$ , where  $N_{nat}$  is

the number of pairs of beads with native interactions in the protein conformation (see SI), and  $L_i$  is the number of residues separating the contact pair  $i$ .

## Results and Discussion

**Thermodynamics of Protein L folding:** The protein in the folded state has one  $\alpha$ -helix ( $\alpha_1$ ) and four  $\beta$ -strands ( $\beta_1 - \beta_4$ ) (Fig. 1A and S1A). Low friction Langevin dynamics simulations performed at different temperatures,  $T$ , ranging from  $300K$  to  $430K$  show that folding occurs in a two-state manner (see SI, Fig. S1). The melting temperature,  $T_M$ , of protein L obtained from the heat capacity,  $C_v$ , plot is  $374.5K$  (Fig. S1), and the value observed in experiments<sup>58</sup> is  $348.5K$ . The difference in  $T_M$  between experiments and simulations can be attributed to the simplified coarse-grained SOP-SC model. At  $T_M$ , the protein transitions between the native basin of attraction (NBA) and the unfolded basin of attraction (UBA) (Fig. S2). Protein L folding thermodynamics and kinetics in the presence of the denaturant *GuHCl* is studied using the molecular transfer model (MTM)<sup>11,43</sup>. In order to compare the denaturant-dependent folding properties of the protein computed from simulations with the experiments, a simulation temperature  $T_S (= 357.7K)$  at which theoretically obtained free energy difference between the NBA and UBA,  $\Delta G_{NU}^{Sim} (= G_N(T_S) - G_U(T_S))$  matches with the experimentally<sup>29</sup> measured value,  $\Delta G_{NU}^{Exp} (= -4.6 \text{ kcal/mole})$ , at  $[GuHCl] = 0M$  is used. This is the only adjustable parameter in the model, which is equivalent to matching the energy scales between the simulations and experiments.

The structural overlap parameter,  $\chi$  (see methods), is used to distinguish between the NBA and UBA protein conformations (Fig. S2). The protein conformations with  $\chi \leq 0.47$  belong to the NBA, and conformations with  $\chi > 0.47$  belong to the UBA (Fig. S2B). The fraction of molecules in NBA,  $f_{NBA}$ , as a function of  $[GuHCl]$  computed from simulations is in quantitative agreement with the experiments<sup>23,29,59</sup> (Fig. 1B). The mid-point  $[GuHCl]$  at which the protein unfolds is  $\approx 2.5M$ . The average radius of gyration,  $\langle R_g \rangle$ , of Protein L as a function of  $[GuHCl]$  is in quantitative agreement with the SAXS experiments<sup>25</sup> (Fig. 1C). The standard deviation of  $R_g$  in the protein unfolded state,  $\sigma_{R_g} \approx 4\text{\AA}$ , indicates that  $R_g$  fluctuates between  $22\text{\AA} \lesssim R_g \lesssim 30\text{\AA}$ . As  $[GuHCl]$  is diluted from  $8M$  to  $4M$ , the  $\langle R_g \rangle$  decreases from  $\approx 26.5\text{\AA}$  to  $\approx 24.7\text{\AA}$  almost

linearly with a slope of  $0.42 \text{ \AA M}^{-1}$ . The experimental data<sup>25</sup> fits equally well with a horizontal line or a line with slope  $0.33 \pm 0.35 \text{ \AA M}^{-1}$ . The average  $R_g$  of the protein conformations in the UBA basin,  $\langle R_g^{UBA} \rangle$ , as a function of  $[GuHCl]$  show that the size of the protein decreases from  $\approx 26.5 \text{ \AA}$  to  $\approx 22.5 \text{ \AA}$  as  $[GuHCl]$  is diluted from 7.5M to 0.25M.

**Denaturant-dependent FRET efficiency:** Average FRET efficiency,  $\langle E \rangle$ , as a function of  $[GuHCl]$  is computed from the Langevin dynamics simulations (Fig. 2A). In FRET experiments<sup>21,22</sup>, the donor (AlexaFluor 488) and acceptor (AlexaFluor 594) dyes are attached near the N and C termini of Protein L. All-atom simulations<sup>60</sup> have shown that the dyes have negligible effect on the size of disordered protein structures. The  $\langle E \rangle$  is calculated using

$$\langle E \rangle = \int_0^L \frac{P(R_{ee})}{1 + \left(\frac{R_{ee}}{R_0}\right)^6} dR_{ee}, \quad (4)$$

where  $P(R_{ee})$  is the end-to-end distance,  $R_{ee}$ , probability distribution function of the protein,  $L(= 248 \text{ \AA})$  is the contour length of the protein, and  $R_0(= 54 \text{ \AA})$  is the Forster radius for the donor-acceptor dyes used in the experiments<sup>21,22</sup>.

The equilibrium  $\langle E \rangle$  transitions from lower values ( $< 0.5$ ) to higher values ( $\approx 0.85$ ) as the protein folds from an unfolded state upon  $[GuHCl]$  dilution (Fig. 2A). The large standard-deviation,  $\sigma_E \approx 0.3$ , for  $[GuHCl] > 2.5M$  (Fig. S3) indicates that the protein in the UBA basin samples conformations with large size fluctuations in agreement with the  $R_g$  data (Fig. 1C). The average FRET efficiency computed for the UBA ensemble,  $\langle E^{UBA} \rangle$ , to study whether the protein collapses in the early stages of folding is in quantitative agreement with the experiments of Eaton *et al.*<sup>22</sup>, where as they are in disagreement with the experiments of Haran *et al.*<sup>21</sup> for the denaturant concentrations  $1M \lesssim [GuHCl] \lesssim 3M$  (Fig. 2A).  $\langle E^{UBA} \rangle$  gradually increases from 0.37 to 0.6 as  $[GuHCl]$  is diluted from 8M to 0.25M pointing to an average decrease in the size of the protein (Fig. 2A).

The average FRET efficiency  $\langle E^{Burst} \rangle$  is also computed from the initial 0.25 *milliseconds* (*ms*) of the Brownian dynamics simulations performed to study the folding kinetics of Protein L. The initial 0.25 *ms* of the folding trajectories are used to check whether the protein decreases in size in the burst-phase of folding when  $[GuHCl]$  is diluted from 7.5M to lower concentrations. The initial unfolded protein conformation to initiate the folding simulations in various  $[GuHCl]$  are obtained from simulations

performed at  $[GuHCl] = 7.5M$ . 20 independent simulations starting from different initial protein conformations are performed for each  $[GuHCl]$ .  $\langle E^{Burst} \rangle$  computed for the early stages of folding is in quantitative agreement with the experiments of Eaton *et al.*<sup>22</sup> for all  $[GuHCl]$ , and deviates from the values obtained from the experiments of Haran *et al.*<sup>21</sup> for the  $[GuHCl]$  range  $1M \lesssim [GuHCl] \lesssim 3M$  (Fig. 2B).  $\langle E^{Burst} \rangle$  increases from 0.38 to 0.53 as  $[GuHCl]$  is diluted from 7.5M to 1.0M signifying that the protein on an average decreases in size. The standard deviation of burst-phase FRET efficiency,  $\sigma_E \approx 0.3$ , show that  $\langle E^{Burst} \rangle$  varies between 0.2 and 0.8 indicating that the protein in the initial stages of folding samples conformations with a significant variation in size (Fig. 2B). The  $R_g$  plot as a function of time shows that it varies in the range  $15\text{\AA} \lesssim R_g \lesssim 30\text{\AA}$  during the initial hundreds of microseconds after folding is initiated for  $[GuHCl] = 1M$  and  $2M$  conditions (Fig. S4).

**FRET overestimates radius of gyration in high  $[GuHCl]$ :** We mimicked the FRET experiments<sup>21,22</sup> to estimate  $\langle R_g^{FRET} \rangle$  from  $\langle E^{Burst} \rangle$ . The Gaussian polymer chain end-to-end probability distribution is given by

$$P(R_{ee}) = 4\pi R_{ee}^2 \left( \frac{3}{2\pi \langle R_{ee}^2 \rangle} \right)^{3/2} \exp \left( -\frac{3R_{ee}^2}{2\langle R_{ee}^2 \rangle} \right). \quad (5)$$

The  $P(R_{ee})$  given by Eq. 5 is used in Eq. 4 to estimate the average end-to-end distance square,  $\langle R_{ee}^2 \rangle$ , from  $\langle E \rangle$ .  $\langle R_g \rangle$  is calculated using the relation<sup>61</sup>,  $\langle R_g \rangle = \sqrt{\langle R_{ee}^2 \rangle / 6}$ .  $\langle R_g^{FRET} \rangle$  values estimated from  $\langle E^{Burst} \rangle$  (Fig. 2B) using equations 4 and 5 at different  $[GuHCl]$  are in near quantitative agreement with the experimentally<sup>22</sup> estimated values (Fig. 3A). On diluting  $[GuHCl]$  from 7.5M to 1M,  $\langle R_g^{FRET} \rangle$  decreases from  $\approx 30\text{\AA}$  to  $\approx 24\text{\AA}$ , nearly a  $6\text{\AA}$  change in the size of the protein, which is in agreement with the experiments<sup>22</sup> of Eaton *et al.* However,  $\langle R_g^{FRET} \rangle$  deviates from the  $\langle R_g^{Burst} \rangle$  values computed directly from the protein conformations obtained from the simulation trajectories (Fig. 3A).  $\langle R_g^{Burst} \rangle$  decreases from  $\approx 26.7\text{\AA}$  to  $\approx 23.8\text{\AA}$ , a decrease of only  $\approx 3\text{\AA}$  upon  $[GuHCl]$  dilution from 7.5M to 1M (Fig. 3A). This shows that FRET overestimates the size of the protein especially in higher  $[GuHCl]$ , and this gives rise to the appearance of pronounced compaction in the dimensions of the protein in the burst-phase as  $[GuHCl]$  is diluted. The standard deviation,  $\sigma_{R_g} \approx 4\text{\AA}$ , of  $\langle R_g^{Burst} \rangle$  shows that at all  $[GuHCl]$ , the protein samples conformations with  $R_g$  varying from  $\approx 22\text{\AA}$  to  $\approx 28\text{\AA}$  (Fig. 3A), and the  $3\text{\AA}$  decrease in  $\langle R_g^{Burst} \rangle$  upon  $[GuHCl]$  dilution



is within the  $\sigma_{R_g}$ . The deviation between  $\langle R_g^{FRET} \rangle$  and  $\langle R_g^{Burst} \rangle$  at high  $[GuHCl]$  was also emphasised in the work of O'Brien *et al.*<sup>62</sup> and the reasons for the deviation are attributed to the use of Gaussian chain  $P(R_{ee})$  to extract  $\langle R_g^{Burst} \rangle$ . The problems associated with the use of Gaussian chain  $P(R_{ee})$  to extract information about protein dimensions is also highlighted in previous other studies<sup>63–66</sup>. The deviation between  $\langle R_g^{FRET} \rangle$  and  $\langle R_g^{Burst} \rangle$  increases with  $[GuHCl]$ , and the reasons for the discrepancy can be understood using the relation  $\langle R_{ee}^2 \rangle = \langle R_{ee} \rangle^2 + \sigma_{R_{ee}}^2$ , where  $\sigma_{R_{ee}}$  is the standard deviation in  $R_{ee}$ .

$\langle R_{ee}^{FRET} \rangle$  and  $\sigma_{R_{ee}}^{FRET}$  estimated from the Gaussian polymer chain  $P(R_{ee})$  to compute  $\langle R_g^{FRET} \rangle$  deviate from the values  $\langle R_g^{Burst} \rangle$  computed directly from the initial 0.25ms of the Protein L folding trajectories (Fig. 3B and C). At high  $[GuHCl](= 7.5M)$ ,  $\langle R_{ee}^{FRET} \rangle$  and  $\sigma_{R_{ee}}^{FRET}$  estimated from the Gaussian chain  $P(R_{ee})$  are 67.3Å and 28.3Å, respectively, which deviate from the  $\langle R_{ee}^{Burst} \rangle$  and  $\sigma_{R_{ee}}^{Burst}$  values 62.7Å and 20.5Å respectively (Fig. 3B), computed directly from the simulations. As a result FRET overestimates  $\langle R_g^{FRET} \rangle \left( = \sqrt{\langle (R_{ee}^{FRET})^2 \rangle / 6} = \sqrt{[\langle R_{ee}^{FRET} \rangle^2 + (\sigma_{R_{ee}}^{FRET})^2] / 6} \right)$  compared to  $\langle R_g^{Burst} \rangle$  in high  $[GuHCl]$  (Fig. 3A). As  $[GuHCl]$  decreases, the deviation between  $\langle R_g^{FRET} \rangle$  and  $\langle R_g^{Burst} \rangle$  decreases (Fig. S5). In low  $[GuHCl](= 1.0M)$ , the  $\langle R_{ee}^{FRET} \rangle$  values computed from the Gaussian chain  $P(R_{ee})$ , and  $\langle R_{ee}^{Burst} \rangle$  computed from simulations are in good agreement, where as the  $\sigma_{R_{ee}}^{FRET}$  values deviate from  $\sigma_{R_{ee}}^{Burst}$  (Fig. 3C). Due to this the deviation between  $\langle R_g^{Burst} \rangle$  and  $\langle R_g^{FRET} \rangle$  is small in low  $[GuHCl]$ , and increases with  $[GuHCl]$  (Fig. 3 and S5).

To conclude, during the burst-phase of folding,  $\langle R_g^{Burst} \rangle$  for Protein L decreases from  $\approx 26.7(\pm 4)\text{Å}$  to  $\approx 23.8(\pm 4)\text{Å}$ , a decrease of  $\approx 3\text{Å}$  upon  $[GuHCl]$  dilution from 7.5M to 1M (Fig. 3A). However, FRET overestimates the size of the protein in high  $[GuHCl](\approx 7.5M)$  due to the application of the Gaussian polymer chain  $P(R_{ee})$  to estimate  $\langle R_g^{FRET} \rangle$ . During the burst-phase  $\langle R_g^{FRET} \rangle$  estimated from FRET decreases from  $\approx 30\text{Å}$  to  $\approx 24\text{Å}$  upon  $[GuHCl]$  dilution from 7.5M to 1M (Fig. 3A). Due to the  $\approx 6\text{Å}$  decrease in  $\langle R_g^{FRET} \rangle$ , FRET experiments<sup>21,22</sup> suggest pronounced compaction in the protein size during the burst-phase of folding.

**Disagreement between FRET and SAXS experiments on Protein L compaction in burst-phase folding:** In simulations, the average radius of gyration in the burst-phase folding,  $\langle R_g^{Burst} \rangle$ , decreased by  $\approx 3\text{Å}$  when  $[GuHCl]$  is diluted from 7.5M to 1M. The  $\approx 3\text{Å}$  decrease in  $\langle R_g^{Burst} \rangle$  is close to statistical uncertainties of the  $R_g$  data ob-

tained from SAXS experiments<sup>25</sup> for Protein L. The  $R_g$  data from SAXS experiments for  $3M \lesssim [GuHCl] \lesssim 7M$  can fit a horizontal line or a line with slope  $0.33 \pm 0.35 \text{ \AA M}^{-1}$  equally well<sup>25</sup>. Using this slope to compute the  $R_g$  change upon  $[GuHCl]$  dilution from  $\approx 7.5M$  to  $1M$  gives a  $R_g$  decrease of  $2.1 \pm 2.3 \text{ \AA}$ . The SAXS experiments<sup>25</sup> report that  $R_g$  of Protein L in the burst phase upon  $[GuHCl]$  dilution to  $1.3M$  and  $0.67M$  are  $\approx 23.5 \pm 2.1 \text{ \AA}$  and  $24.9 \pm 1.12 \text{ \AA}$ , respectively which are statistically not different from the value  $23.7 \pm 0.4$  at  $[GuHCl] = 4.0M$ . In the simulations,  $\langle R_g^{Burst} \rangle$  at  $[GuHCl] = 4M$  and  $1M$  are  $25.1 \pm 4.0 \text{ \AA}$  and  $23.8 \pm 3.9 \text{ \AA}$ , respectively, a difference of  $\approx 1.3 \text{ \AA}$ , which is well within  $\sigma_{R_g} \approx 4 \text{ \AA}$ . This analysis similar to the SAXS analysis leads to the conclusion of minimal Protein L compaction within statistical uncertainties on  $[GuHCl]$  dilution in agreement with the SAXS experiments<sup>25</sup>, and disagreement with the FRET experiments<sup>21,22</sup>.

Recent FRET experiments<sup>67</sup> on polyethylene glycol (PEG) showed that FRET efficiency decreased as  $[GuHCl]$  is increased when hydrophilic PEG is unlikely to expand on increasing  $[GuHCl]$ . This led to questions about the interpretation of the FRET data to study protein collapse in low  $[GuHCl]$ . We find that the computed variation in  $\langle E \rangle$  and  $\langle R_g \rangle$  as a function of  $[GuHCl]$  for Protein L to be in quantitative agreement with at least one of the FRET experiments<sup>22</sup> (Fig. 2) and also in agreement with SAXS experiments<sup>25</sup> within the statistical uncertainties (Fig. 1C and 3A). The results points to the use of Gaussian polymer chain statistics to extract  $R_g$  from FRET efficiency data to be the cause for the discrepancy between the SAXS and FRET experiments in estimating  $R_g$ .

**The coil-globule transition in Protein L is concomitant with the folding transition:** In polymers the ratio of the radius of gyration to the hydrodynamic radius,  $R_g/R_h$ , can point to the coil-globule collapse transition. The  $R_g/R_h$  ratio for a polymer in a good solvent<sup>68</sup> is  $\approx 1.56$ , where as the ratio in a poor solvent<sup>61</sup> is  $\approx 0.77$ . We used the Kirkwood-Riseman approximation<sup>69</sup> to compute the hydrodynamic radius of the protein, which is given by  $R_h = (1/2N^2) \sum_{i \neq j} 1/|\vec{r}_i - \vec{r}_j|$ , where  $N$  is the number of beads in the coarse-grained protein,  $\vec{r}_i$  and  $\vec{r}_j$  are the position vectors of beads  $i$  and  $j$ . The  $R_g/R_h$  ratio for the burst-phase folding decreases from  $\approx 1.31$  to  $\approx 1.28$  as  $[GuHCl]$  is diluted from  $7.5M$  to  $1M$  indicating that this is not a coil-globule transition observed in polymers (Fig. 4). The single domain proteins which are finite in size compared to polymers are predicted to have a near overlap of the collapse and folding transition temperatures<sup>12,13</sup>. In

agreement, the equilibrium ratio of  $R_g/R_h$  decreases from  $\approx 1.3$  to  $\approx 0.98$  as the folding transition occurs (Fig. 4). The  $R_g/R_h$  ratio does not approach 0.77 as the protein folds because the Kirkwood-Riseman approximation<sup>69</sup> used to compute  $R_h$  does not hold for the protein in the folded state as it assumes all the beads are equally bathed by the solvent. The absence of coil-globule transition in the burst-phase of protein L folding does not imply that the collapse transition or significant protein compaction is universally absent in the burst-phase folding of all single domain proteins. Both FRET and SAXS experiments agree that the protein Monellin<sup>18–20</sup> shows compaction during the burst-phase of folding. Although both the experimental techniques observe compaction in the case of Cytochrome *c*<sup>14–17</sup>, the FRET experiments show that this compaction, a sub-100 $\mu$ s event, is barrier limited and it is due to the formation of marginally stable partially folded structures<sup>14</sup>. Experiments<sup>8</sup> show that for the protein CyclophilinA, the  $R_g/R_h$  ratio decreases from a value between 1.1-1.2 to a value between 0.9-1.0 as *[GuHCl]* is diluted from 8M to 0M indicating a coil-globule transition in the burst-phase of folding.

**Transition State Ensemble (TSE):** The transition state ensemble of Protein L at the melting temperature,  $T_M$ , is identified using the  $P_{fold}$  analysis<sup>44</sup> (see SI for details). 12 out of 108 putative transition state structures (TSE) which satisfy the condition,  $0.4 < P_{fold} < 0.6$  are labeled as TSE (Fig. S6). The transition state structures (TSE) are globular, extensive and homogenous, with most of the secondary and tertiary contacts formed (Fig. 5). The  $\Psi$ -analysis experiments<sup>30</sup> predict that TSE contains all the four  $\beta$ -sheet strands ( $\beta_1 - \beta_4$ ). The TSE from simulations show that both the *N* and *C*-termini hairpins  $\beta_1\beta_2$  and  $\beta_3\beta_4$ , and the contacts between the strands  $\beta_1\beta_4$  are present in the structures in agreement with the  $\Psi$ -analysis experiments<sup>30</sup> (Fig. 5B).

The  $\Psi$ -analysis experiments on two residue pairs, K28-E32 and A35-T39, present in the helical region of the protein gave  $\Psi$ -values 0.26 and  $\ll 0$ , respectively, indicating that contacts between these pairs of residues is largely absent, and concluded that helix  $\alpha_1$  is mostly not present in the TSE<sup>30</sup>. The contact map of the TSE obtained from the simulations show that the side chains of the residue pairs K28-E32 and A35-T39 form contacts with a probability of 0.41 and 0.08, respectively. The simulations further indicate that a cluster of residues between S31 and A37 present approximately at the center of the helix containing 3 Ala residues (A33, A35 and A37) can form stable contacts in the TSE (Fig. 5B).

The  $\Psi$ -analysis experiments<sup>31,70</sup> predict a relationship between the relative contact order, RCO (see methods), of the native protein topology and TSE,  $RCO^{TSE} \approx 0.7RCO^{Native}$ , which shows the extent of long-range contacts present in the TSE compared to the native-state. The TSE structures extracted from the simulations show  $RCO^{TSE}/RCO^{Native} = 0.77$ , which is in reasonable agreement with the value of 0.75 estimated from  $\Psi$ -analysis experiments<sup>30</sup>. The simulations using the coarse-grained protein model support the basic topology of the TSE structures predicted by the  $\Psi$ -analysis experiments.

The folding simulations of only the C-terminal hairpin ( $\beta_3\beta_4$ ) using atomistic models predicted the presence of non-native contacts, a 2 amino acid register shift, in the TSE<sup>30</sup>. We do not observe this 2 amino acid register shift in the C-terminal hairpin because the SOP-SC model includes only native-interactions. The predicted TSE is only in partial agreement with the  $\Phi$ -analysis experiments<sup>27-29,71</sup> which predicted a polarised structure with only  $\beta_1\beta_2$  hairpin. The results support the hypothesis that folding pathways and TSE of single domain proteins are influenced by the topology of the folded structure in agreement with the experiments<sup>45</sup>.

**Concluding Remarks:** In summary, we have studied Protein L folding in the presence of the denaturant Guanidine Hydrochloride using the SOP-SC coarse-grained model and molecular dynamics simulations. The effect of  $[GuHCl]$  on the protein is taken into account using the molecular transfer model<sup>11,43</sup>. The study mainly focussed on whether there is a coil-globule collapse transition in the burst-phase of folding after the denaturant concentration is diluted to lower values. The main findings of this study is the coil-globule transition in Protein L is concomitant with the folding transition. It is not observed during the burst phase. The FRET experiments overestimate the  $R_g$  of the protein at high  $[GuHCl]$  concentrations owing to the use of the Gaussian polymer chain end-to-end distribution function to extract  $R_g$  from FRET efficiency. As a result, in the burst-phase of folding, FRET observes pronounced compaction in the size of the protein as  $[GuHCl]$  is diluted. The actual decrease in the size of the protein ( $\approx 3\text{\AA}$ ) observed during the burst-phase is close to statistical uncertainties of the  $R_g$  data measured from SAXS experiments<sup>25</sup>, and these experiments conclude that there is no collapse leading to a discrepancy with the FRET experiments.

It is highly desirable to formulate a method to accurately extract the distances between the donor and acceptor dyes used in the FRET experiments. However, it is a non-trivial

inverse problem as we seek to accurately extract a probability distribution of the distances between the dyes from the average FRET efficiency measured in experiments, especially in cases like Protein L where the compaction in protein dimensions is small on denaturant dilution. Previous studies<sup>62</sup> have shown that even other polymer models such as the self-avoiding chain or the worm-like-chain model are also not very accurate quantitatively to predict the small subtle changes in the protein dimensions.

The results presented in this manuscript clearly point out the aspects of the Gaussian chain model, which leads to over estimating the size of the protein when used to analyse the FRET data. The results show that the Gaussian chain model fails in accurately capturing the width of the protein end-to-end probability distribution, which is essential to compute the radius of gyration. For any method to be quantitatively accurate it should capture the peak position as well as the width of the probability distribution accurately, and this is a challenging task because we need to estimate probability distribution from an average value, and also the method should be reliable enough to work on proteins with different amino acid composition and native folds.

To check the accuracy of the distance between the dyes extracted from the FRET efficiency data using the Gaussian polymer model assumption, a self-consistency check can be performed to see if the assumption is valid or not for the protein under study<sup>62</sup>. If the dyes are attached at locations  $i$  and  $j$  in the protein, and  $\langle R_{ij}^2 \rangle$  is the average distance square extracted from FRET efficiency, and similarly if  $\langle R_{kl}^2 \rangle$  is the average distance square extracted from FRET efficiency with dyes at positions  $k$  and  $l$ , then the relation  $\langle R_{ij}^2 \rangle / \langle R_{kl}^2 \rangle = |j - i| / |l - k|$  should hold if the protein behaves as a Gaussian chain. If the relation is not satisfied, then one should be cautious in quantitatively inferring results about the protein dimensions assuming that the protein in the unfolded state behaves as a Gaussian chain.

The magnitude of protein compaction in the burst-phase folding of single domain proteins upon denaturant dilution is not uniform, and it should depend on protein length, sequence and composition of amino acids. For example SAXS experiments on Protein L<sup>25</sup> and Ubiquitin<sup>72</sup> infer no compaction in the protein dimensions in the burst-phase, while experiments on Cytochrome  $c$ <sup>14</sup> and Monellin<sup>20</sup> observe compaction. The key features in the single domain proteins responsible for compaction in protein dimensions on denaturant dilution needs to be identified. In addition to temperature and denaturants, force can also be

used to unfold proteins and study protein folding. Experiments<sup>73</sup> and simulations<sup>74,75</sup> show that a protein unfolded by force when allowed to refold in the presence of lower quenching forces undergoes a rapid compaction in the initial stages of folding. This compaction of the protein is driven by entropy because the protein in the stretched state is in a low entropic state and upon force quench undergoes rapid compaction in the first stage of folding until entropy is maximised<sup>75</sup>. The extent of protein compaction in the initial stages of folding also depends on the experimental probes used to study protein folding.

The transition state structures inferred from the  $P_{fold}$  analysis are globular and extensive with both the C and N-termini hairpins  $\beta_1\beta_2$  and  $\beta_3\beta_4$ , and interactions between the strands  $\beta_1\beta_4$ . These results are in agreement with the  $\Psi$ -analysis experiments<sup>30</sup> and support the hypothesis that for single domain globular proteins the transition state structures depend on the protein native-state topology and are not stabilised by local interactions alone.

**Acknowledgement:** GR acknowledges startup grant from Indian Institute of Science-Bangalore, and funding from Nano mission, Department of Science and Technology, India. Hiranmay Maity acknowledges research fellowship from Indian Institute of Science-Bangalore.

**Supporting Information:** Description of the simulation methods; Table S1; Figures S1-S6. This material is available free of charge via the Internet at <http://pubs.acs.org>

- 
- \* Electronic address: greddy@sscu.iisc.ernet.in
- <sup>1</sup> G. Ziv, D. Thirumalai, and G. Haran. Collapse transition in proteins. *Phys. Chem. Chem. Phys.*, 11:83–93, 2009.
- <sup>2</sup> Tobin R. Sosnick and Doug Barrick. The folding of single domain proteins - have we reached a consensus? *Curr. Opin. Struct. Biol.*, 21(1):12–24, 2011.
- <sup>3</sup> Gilad Haran. How, when and why proteins collapse: the relation to folding. *Curr. Opin. Struct. Biol.*, 22(1):14–20, 2012.
- <sup>4</sup> D Thirumalai, Zhenxing Liu, Edward P O’Brien, and Govardhan Reddy. Protein folding: From theory to practice. *Curr. Opin. Struct. Biol.*, 23(1):22–29, 2013.
- <sup>5</sup> Jayant B. Udgaonkar. Polypeptide chain collapse and protein folding. *Arch. Biochem. Biophys.*, 531(1-2, SI):24–33, 2013.
- <sup>6</sup> C Tanford, K Kawahara, and S Lapanje. Proteins in 6M Guanidine Hydrochloride - Demonstration of random coil behavior. *J. Biol. Chem.*, 241(8):1921–&, 1966.
- <sup>7</sup> J.E. Kohn, I.S. Millett, J. Jacob, B. Zagrovic, T.M. Dillon, N. Cingel, R.S. Dothager, S. Seifert, P. Thiagarajan, T.R. Sosnick, M.Z. Hasan, V.S. Pande, I. Ruczinski, S. Doniach, and K.W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl Acad Sci USA*, 101(34):12491–12496, 2004.
- <sup>8</sup> H Hofmann, A Soranno, A Borgia, K Gast, D Nettels, and B Schuler. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single molecule spectroscopy. *Proc. Natl. acad. sci.*, 109:16155–16160, 2012.
- <sup>9</sup> J D Bryngelson and P G Wolynes. A Simple Statistical Field-Theory of Heteropolymer collapse with application to protein folding. *Biopolymers*, 30(1-2):177–188, 1990.
- <sup>10</sup> H S Chan and K A Dill. Polymer principles in protein-structure and stability. *Annu. Rev. Biophys. Biophys. Chem.*, 20:447–490, 1991.
- <sup>11</sup> E. P. O’Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. USA*, 105:13403–13408, 2008.
- <sup>12</sup> C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad Sci USA*, 90(13):6369–6372, 1993.

- <sup>13</sup> M S Li, D K Klimov, and D Thirumalai. Finite size effects on thermal denaturation of globular proteins. *Phys. Rev. Lett.*, 93(26), 2004.
- <sup>14</sup> Sagar V. Kathuria, Can Kayatekin, Raul Barrea, Elena Kondrashkina, Rita Graceffa, Liang Guo, R. Paul Nobrega, Srinivas Chakravarthy, C. Robert Matthews, Thomas C. Irving, and Osman Bilsel. Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS. *J. Mol. Biol.*, 426(9):1980–1994, 2014.
- <sup>15</sup> Lisa J. Lapidus, Shuhuai Yao, Kimberly S. McGarrity, David E. Hertzog, Emily Tubman, and Olgica Bakajin. Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. *Biophys. J.*, 93(1):218–224, 2007.
- <sup>16</sup> S J Hagen and W A Eaton. Two-state expansion and collapse of a polypeptide. *J. Mol. Biol.*, 301(4):1019–1027, 2000.
- <sup>17</sup> M C R Shastry, J M Sauder, and H Roder. Kinetic and structural analysis of submillisecond folding events in cytochrome c. *Accounts Chem. Res.*, 31(11):717–725, 1998.
- <sup>18</sup> Rama Reddy Goluguri and Jayant B. Udgaonkar. Rise of the Helix from a Collapsed Globule during the Folding of Monellin. *Biochemistry*, 54(34):5356–5365, 2015.
- <sup>19</sup> Santosh Kumar Jha and Jayant B. Udgaonkar. Direct evidence for a dry molten globule intermediate during the unfolding of a small protein. *Proc. Natl. Acad. Sci. U. S. A.*, 106(30):12289–12294, 2009.
- <sup>20</sup> T Kimura, T Uzawa, K Ishimori, I Morishima, S Takahashi, T Konno, S Akiyama, and T Fujisawa. Specific collapse followed by slow hydrogen-bond formation of beta-sheet in the folding of single-chain monellin. *Proc. Natl. Acad. Sci.*, 102(8):2748–2753, 2005.
- <sup>21</sup> E. Sherman and G. Haran. Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. USA*, 103:11539–11543, 2006.
- <sup>22</sup> K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. USA*, 104:1528–1533, 2007.
- <sup>23</sup> Steven A. Waldauer, Olgica Bakajin, Terry Ball, Yujie Chen, Stephen J. DeCamp, Michaela Kopka, Marcus Jaeger, Vijay R. Singh, William J. Wedemeyer, Shimon Weiss, Shuhuai Yao, and Lisa J. Lapidus. Ruggedness in the folding landscape of protein L. *HFSP J.*, 2(6):388–395, 2008.
- <sup>24</sup> K W Plaxco, I S Millett, D J Segel, S Doniach, and D Baker. Chain collapse can occur



- concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.*, 6(6):554–556, 1999.
- <sup>25</sup> Tae Yeon Yoo, Steve P. Meisburger, James Hinshaw, Lois Pollack, Gilad Haran, Tobin R. Sosnick, and Kevin Plaxco. Small-Angle X-ray Scattering and Single-Molecule FRET Spectroscopy Produce Highly Divergent Views of the Low-Denaturant Unfolded State. *J. Mol. Biol.*, 418(3-4):226–236, 2012.
- <sup>26</sup> M L Scalley, Q Yi, H D Gu, A McCormack, J R Yates, and D Baker. Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry*, 36(11):3373–3382, 1997.
- <sup>27</sup> H D Gu, D Kim, and D Baker. Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein. *J. Mol. Biol.*, 274(4):588–596, 1997.
- <sup>28</sup> D E Kim, Q Yi, S T Gladwin, J M Goldberg, and D Baker. The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.*, 284(3):807–815, DEC 4 1998.
- <sup>29</sup> D E Kim, C Fisher, and D Baker. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.*, 298(5):971–984, 2000.
- <sup>30</sup> Tae Yeon Yoo, Aashish Adhikari, Zhen Xia, Tien Huynh, Karl F. Freed, Ruhong Zhou, and Tobin R. Sosnick. The Folding Transition State of Protein L Is Extensive with Nonnative Interactions (and Not Small and Polarized). *J. Mol. Biol.*, 420(3):220–234, 2012.
- <sup>31</sup> Michael C. Baxa, Karl F. Freed, and Tobin R. Sosnick. Quantifying the structural requirements of the folding transition state of protein A and other systems. *J. Mol. Biol.*, 381(5):1362–1381, 2008.
- <sup>32</sup> Athi N. Naganathan and Victor Munoz. Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl. Acad. Sci. U. S. A.*, 107(19):8611–8616, 2010.
- <sup>33</sup> N Koga and S Takada. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model. *J. Mol. Biol.*, 313(1):171–180, 2001.
- <sup>34</sup> J. Karanicolas and C. L. Brooks III. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.*, 11:2351–2361, 2002.
- <sup>35</sup> C Clementi, A E Garcia, and J N Onuchic. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. *J. Mol. Biol.*, 326(3):933–954, 2003.
- <sup>36</sup> M R Ejtehadi, S P Avall, and S S Plotkin. Three-body interactions improve the prediction of

- rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. U. S. A.*, 101(42):15088–15093, 2004.
- <sup>37</sup> S Brown and T Head-Gordon. Intermediates and the folding of proteins L and G. *Protein Sci.*, 13(4):958–970, 2004.
- <sup>38</sup> Qingwu Yang and Sing-Hoi Sze. Predicting protein folding pathways at the mesoscopic level based on native interactions between secondary structure elements. *BMC Bioinformatics*, 9, 2008.
- <sup>39</sup> Vincent A. Voelz, Vijay R. Singh, William J. Wedemeyer, Lisa J. Lapidus, and Vijay S. Pande. Unfolded-State Dynamics and Structure of Protein L Characterized by Simulation and Experiment. *J. Am. Chem. Soc.*, 132(13):4702–4709, 2010.
- <sup>40</sup> Tao Chen and Hue Sun Chan. Effects of desolvation barriers and sidechains on local-nonlocal coupling and chevron behaviors in coarse-grained models of protein folding. *Phys. Chem. Chem. Phys.*, 16(14):6460–6479, 2014.
- <sup>41</sup> C. B. Hyeon, R. I. Dima, and D. Thirumalai. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure*, 14:1633–1645, 2006.
- <sup>42</sup> Zhenxing Liu, Govardhan Reddy, Edward P O’Brien, and D Thirumalai. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. USA*, 108(19):7787–7792, 2011.
- <sup>43</sup> Zhenxing Liu, Govardhan Reddy, and D. Thirumalai. Theory of the Molecular Transfer Model for Proteins with Applications to the Folding of the src-SH3 Domain. *J. Phys. Chem. B*, 116(23):6707–6716, 2012.
- <sup>44</sup> R Du, V S Pande, A Y Grosberg, T Tanaka, and E S Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.
- <sup>45</sup> Michael C. Baxa, Wookyoung Yu, Aashish N. Adhikari, Liang Ge, Zhen Xia, Ruhong Zhou, Karl F. Freed, and Tobin R. Sosnick. Even with nonnative interactions, the updated folding transition states of the homologs Proteins G & L are extensive and similar. *Proc. Natl. Acad. Sci. U. S. A.*, 112(27):8302–8307, 2015.
- <sup>46</sup> J W O’Neill, D E Kim, D Baker, and K Y J Zhang. Structures of the B1 domain of protein L from *Peptostreptococcus magnus* with a tyrosine to tryptophan substitution. *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, 57(4):480–487, 2001.
- <sup>47</sup> M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: Choice of reference

- states and sensitivity of predicted native states to variations in the interaction schemes. *Prot. Sci.*, 8:361–369, 1999.
- <sup>48</sup> Govardhan Reddy and D Thirumalai. Dissecting Ubiquitin folding using the Self-Organized Polymer model. *J. Phys. Chem. B*, 119(34):11358–11370, 2015.
- <sup>49</sup> Govardhan Reddy, Zhenxing Liu, and D. Thirumalai. Denaturant-dependent folding of GFP. *Proc. Natl. Acad. Sci. USA*, 109:17832–17838, 2012.
- <sup>50</sup> M. Auton and D. W. Bolen. Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. *Biochemistry*, 43:1329–1342, 2004.
- <sup>51</sup> E. P. O’Brien, B. R. Brooks, and D. Thirumalai. Molecular Origin of Constant m-Values, Denatured State Collapse, and Residue-Dependent Transition Midpoints in Globular Proteins. *Biochemistry*, 48:3743–3754, 2009.
- <sup>52</sup> T. Veitshans, D. Klimov, and D. Thirumalai. Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold Des*, 2(1):1–22, 1997.
- <sup>53</sup> S Kumar, J M Rosenberg, D Bouzida, R H Swendsen, and P A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. the method. *J. Comput. Chem.*, 13:1011–1021, 1992.
- <sup>54</sup> D. L. Ermak and J. A. Mccammon. Brownian dynamics with hydrodynamic interactions. *J Chem Phys*, 69(4):1352–1360, 1978.
- <sup>55</sup> Z. Guo and D. Thirumalai. Kinetics and thermodynamics of folding of a de novo designed four helix bundle. *J. Mol. Biol.*, 263:323–343, 1996.
- <sup>56</sup> K W Plaxco, K T Simons, and D Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277(4):985–994, 1998.
- <sup>57</sup> D K Klimov and D Thirumalai. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.*, 282(2):471–492, 1998.
- <sup>58</sup> B Gillespie and K W Plaxco. Nonglassy kinetics in the folding of a simple single-domain protein. *Proc. Natl. Acad. Sci. U. S. A.*, 97(22):12014–12019, 2000.
- <sup>59</sup> Eilon Sherman, Anna Itkin, Yosef Yehuda Kuttner, Elizabeth Rhoades, Dan Amir, Elisha Haas, and Gilad Haran. Using fluorescence correlation spectroscopy to study conformational changes in denatured proteins. *Biophys. J.*, 94(12):4819–4827, 2008.
- <sup>60</sup> Guel H. Zerze, Robert B. Best, and Jeetain Mittal. Modest Influence of FRET Chromophores

- on the Properties of Unfolded Proteins. *Biophys. J.*, 107(7):1654–1660, 2014.
- <sup>61</sup> Michael Rubinstein and Ralph H Colby. *Polymer physics*. OUP Oxford, 2003.
- <sup>62</sup> E. P. O’Brien, G. Morrison, B. R. Brooks, and D. Thirumalai. How accurate are polymer models in the analysis of Forster resonance energy transfer experiments on proteins? *J. Chem. Phys.*, 130:124903, 2009.
- <sup>63</sup> Kalyan K Sinha and Jayant B Udgaonkar. Dissecting the non-specific and specific components of the initial folding reaction of barstar by multi-site fret measurements. *J. Mol. Biol.*, 370(2):385–405, 2007.
- <sup>64</sup> T A Laurence, X X Kong, M Jager, and S Weiss. Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 102(48):17348–17353, 2005.
- <sup>65</sup> David P Goldenberg. Computational simulation of the statistical properties of unfolded proteins. *J. Mol. Biol.*, 326(5):1615–1633, 2003.
- <sup>66</sup> Huan-Xiang Zhou. Dimensions of denatured protein chains from hydrodynamic data. *J. Phys. Chem. B*, 106(22):5769–5775, 2002.
- <sup>67</sup> Herschel M. Watkins, Anna J. Simon, Tobin R. Sosnick, Everett A. Lipman, Rex P. Hjelm, and Kevin W. Plaxco. Random coil negative control reproduces the discrepancy between scattering and FRET measurements of denatured protein dimensions. *Proc. Natl. Acad. Sci. U. S. A.*, 112(21):6631–6636, 2015.
- <sup>68</sup> Y Oono and M Kohmoto. Renormalization-group thoery of transport-properties of polymer-solutions .1. Dilute-solutions. *J. Chem. Phys.*, 78(1):520–528, 1983.
- <sup>69</sup> J G Kirkwood and J Riseman. The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J. Chem. Phys.*, 16(6):565–573, 1948.
- <sup>70</sup> Adarsh D. Pandit, Abhishek Jha, Karl F. Freed, and Tobin R. Sosnick. Small proteins fold through transition states with native-like topologies. *J. Mol. Biol.*, 361(4):755–770, 2006.
- <sup>71</sup> M L Scalley, Q Yi, H D Gu, A McCormack, J R Yates, and D Baker. Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry*, 36(11):3373–3382, 1997.
- <sup>72</sup> J Jacob, B Krantz, R S Dothager, P Thiyagarajan, and T R Sosnick. Early collapse is not an obligate step in protein folding. *J. Mol. Biol.*, 338(2):369–382, 2004.
- <sup>73</sup> JM Fernandez and HB Li. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, 2004.

- <sup>74</sup> RB Best and G Hummer. Comment on “Force-clamp spectroscopy monitors the folding trajectory of a single protein”. *Science*, 308(5721):498b, APR 22 2005.
- <sup>75</sup> Changbong Hyeon, Greg Morrison, David L. Pincus, and D. Thirumalai. Refolding dynamics of stretched biopolymers upon force quench. *Proc. Natl. Acad. Sci. U. S. A.*, 106(48):20288–20293, 2009.

# Figures

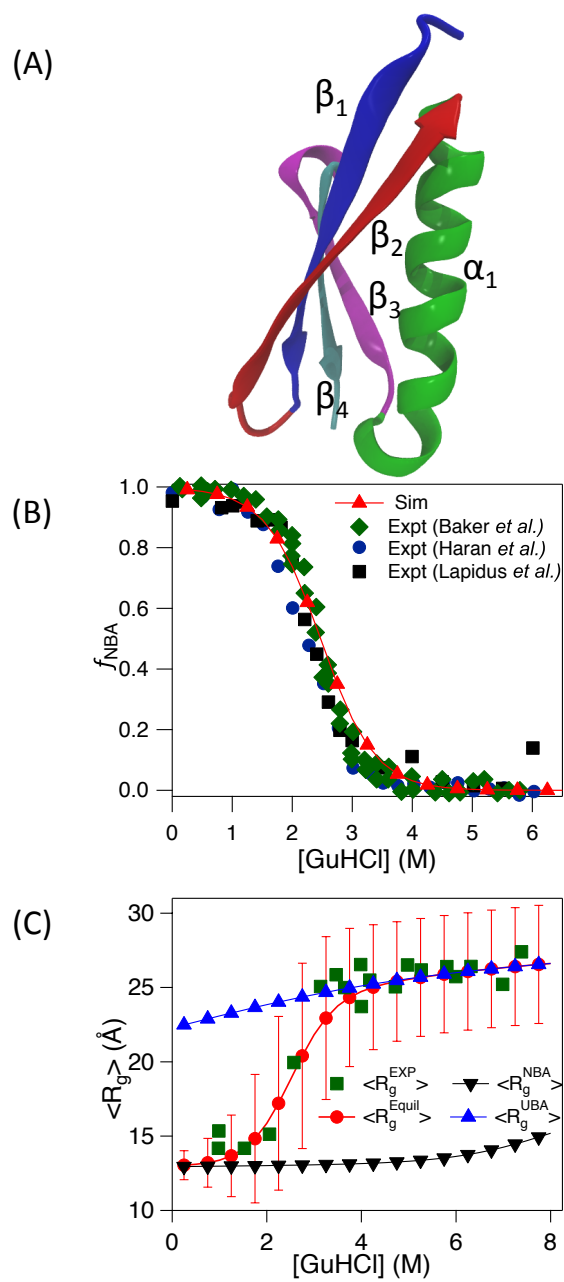


FIG. 1: (A) Crystal structure of Protein L (PDB ID: 1HZ6). The  $\alpha$ -helix is in green ( $\alpha_1$ ), and the four  $\beta$ -strands are in blue ( $\beta_1$ ), red ( $\beta_2$ ), magenta ( $\beta_3$ ), and cyan ( $\beta_4$ ). (B) The fraction of the protein in the native basin of attraction,  $f_{NBA}$ , as a function of  $[GuHCl]$ . Data in red triangles is from simulations. Data in blue circles, black squares and green diamonds are from the experiments of Haran *et al.*<sup>59</sup>, Lapidus *et al.*<sup>23</sup> and Baker *et al.*<sup>29</sup> respectively. (C) The radius of gyration,  $R_g$  as a function of  $[GuHCl]$ . Data in red circles and green squares is from simulations and experiments<sup>25</sup>, respectively.  $\langle R_g \rangle$  of UBA and NBA basins computed from simulations are shown in blue triangles and black inverted triangles, respectively.

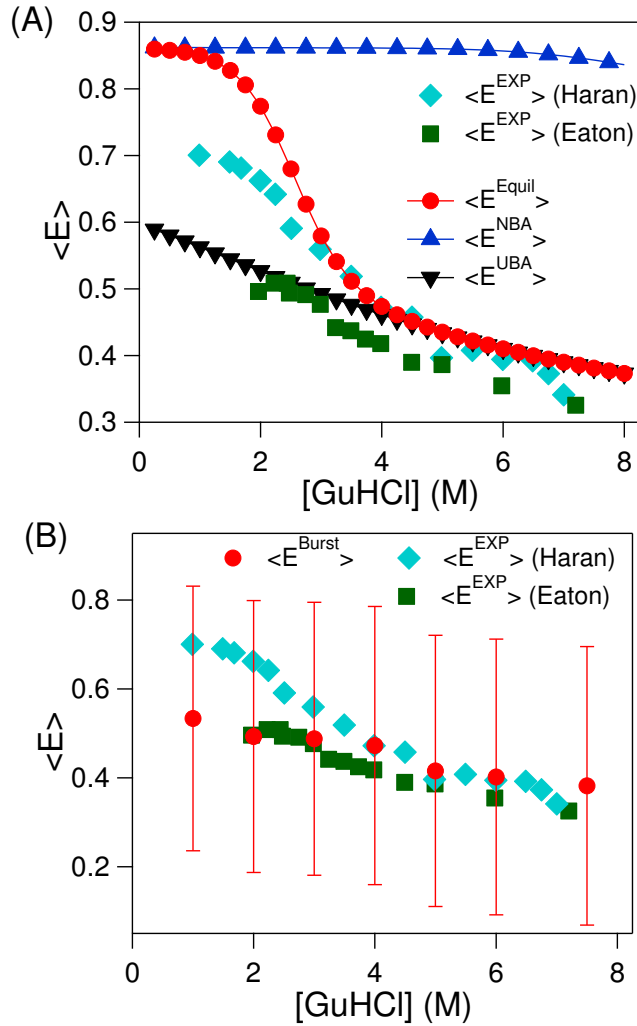


FIG. 2: (A) Equilibrium FRET efficiency,  $\langle E^{Equil} \rangle$ , as a function of  $[GuHCl]$  is in red circles. Experimental data are shown in green squares<sup>22</sup> and cyan diamonds<sup>21</sup>.  $\langle E \rangle$  for the protein conformations in the NBA and UBA basins are shown in blue triangles and black inverted triangles, respectively. (B)  $\langle E^{Burst} \rangle$  shown in red circles is computed from the initial 0.25 *ms* of Protein L Brownian dynamics folding trajectories at  $T = 357.7K$  in various  $[GuHCl]$ . Data in green squares and cyan diamonds is the same as in (A).

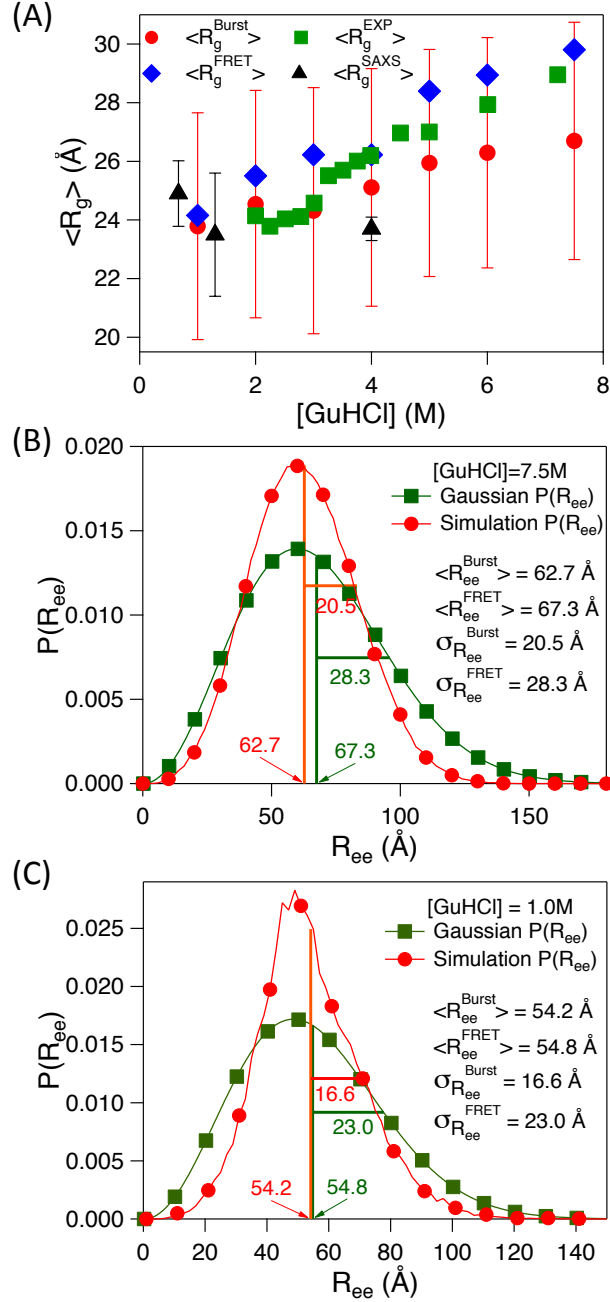




FIG. 3: (A)  $\langle R_g^{FRET} \rangle$  estimated from  $\langle E^{Burst} \rangle$  is in blue diamonds.  $\langle R_g^{Burst} \rangle$  computed from the initial  $0.25ms$  of the Protein L folding trajectories is in red circles. Data in green squares and black triangles is from FRET<sup>22</sup> and SAXS<sup>25</sup> experiments, respectively. (B) The end-to-end distance,  $R_{ee}$ , probability distribution function  $P(R_{ee})$  during the burst phase (initial  $0.25ms$ ) of protein L folding at  $T = 357.7K$  and  $[GuHCl] = 7.5M$  is in red circles.  $P(R_{ee})$  estimated from  $\langle E^{Burst} \rangle$  and Guassian polymer chain statistics in green squares. (C) same as in (B) except that  $[GuHCl] = 1.0M$ .

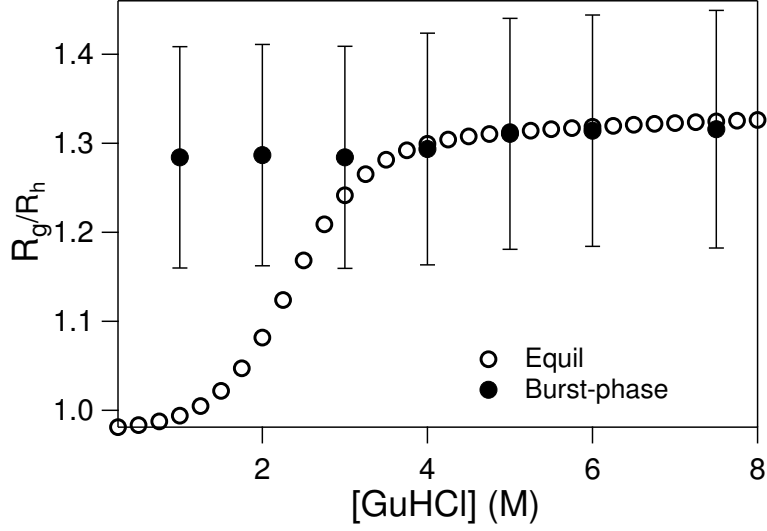


FIG. 4: The ratio of the radius of gyration to the hydrodynamic radius,  $R_g/R_h$ , for the burst phase (solid circles) and equilibrium (empty circles) conditions.

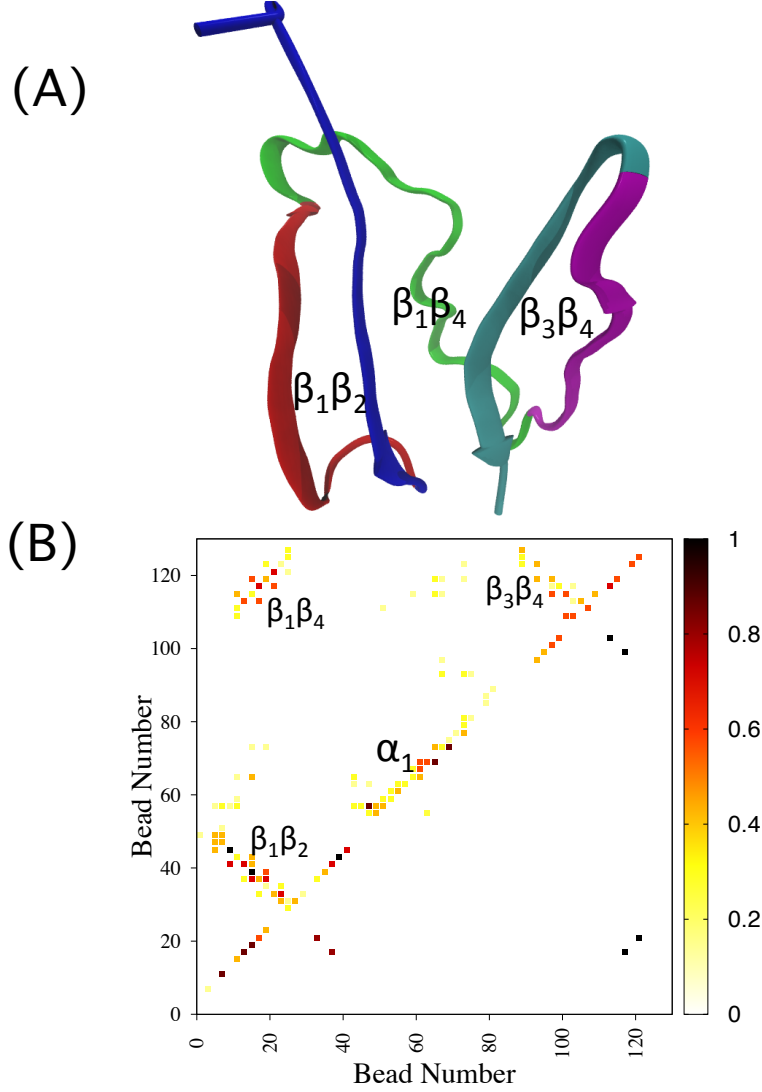


FIG. 5: (A) Representative transition state structure from the transition state ensemble obtained using the  $P_{fold}$  analysis. (B). The contact map of the the transition state ensemble is shown in the upper half of the diagonal. The experimental<sup>30</sup>  $\Psi$ -values for the transition state structure are show in the lower half of the diagonal. The  $\Psi$ -values for the  $\alpha$ -helix residue pairs K28-E32 and A35-T39 from experiments<sup>30</sup> are 0.26 and  $\ll 0$ , respectively. The  $\Psi$ -value used for the residue pair A35-T39 is 0 in the plot. The small  $\Psi$ -values for  $\alpha$ -helix are not visible in the plot.

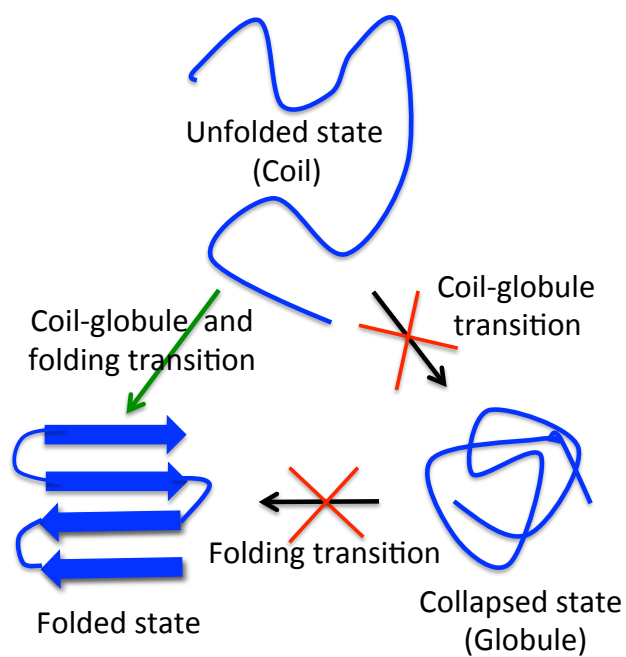


FIG. 6: Table of Contents (TOC) figure

# Supporting Information for “Folding of Protein L with implications for collapse in the denatured state ensemble”

Hiranmay Maity and Govardhan Reddy\*

*Solid State and Structural Chemistry Unit,  
Indian Institute of Science, Bangalore, Karnataka, India 560012*

---

\*Electronic address: greddy@sscu.iisc.ernet.in

## SI Text

### Self Organized Polymer-Side Chain (SOP-SC) model for Protein L:

We used the SOP-SC (self-organized polymer-side chain) model[1, 2] in which each amino acid residue is represented by two beads. One bead is at the  $C_\alpha$  position representing the backbone atoms, and the other bead is at the center of mass of the side chain representing the side chain atoms. The number of residues in Protein L,  $N_{res} = 64$ . The effective energy of a protein conformation in the SOP-SC model is a sum of bonded and non-bonded interactions. The bonded interactions,  $E_B$ , are present between a pair of connected beads. The non-bonded interactions are a sum of native,  $E_{NB}^N$ , and non-native,  $E_{NB}^{NN}$ , interactions. The native interactions for protein L are identified using the crystal structure[3] (Protein Data Bank ID: 1HZ6) (Fig 1A), and they are present between a pair of beads separated by at least 3 bonds, and if the distance between them in the crystal structure is less than  $R_c$  (Table S1).

The coarse-grained force-field in the SOP-SC model for a protein conformation represented by the co-ordinates,  $\{\mathbf{r}\}$ , in the absence of denaturants,  $[C] = 0$ , is

$$E_{CG}(\{\mathbf{r}\}, 0) = E_B + E_{NB}^N + E_{NB}^{NN}. \quad (S1)$$

The bonded interaction energy,  $E_B$ , for all pairs of bonded beads is modelled by finite extensible nonlinear elastic (FENE) potential,

$$E_B = - \sum_{i=1}^{N_B} \frac{k}{2} R_0^2 \log \left( 1 - \frac{(r_i - r_{cry,i})^2}{R_0^2} \right), \quad (S2)$$

where  $N_B (= 127)$  is the total number of pairs of bonds in the SOP-SC model of the protein. The values of  $k$  and  $R_0$  are listed in table S1. Non-bonded native interaction energy,  $E_{NB}^N$ , is modelled by Lennard-Jones type of potential energy and is given by

$$\begin{aligned} E_{NB}^N = & \sum_{i=1}^{N_N^{bb}} \epsilon_h^{bb} \left[ \left( \frac{r_{cry,i}}{r_i} \right)^{12} - 2 \left( \frac{r_{cry,i}}{r_i} \right)^6 \right] + \sum_{i=1}^{N_N^{bs}} \epsilon_h^{bs} \left[ \left( \frac{r_{cry,i}}{r_i} \right)^{12} - 2 \left( \frac{r_{cry,i}}{r_i} \right)^6 \right] \\ & + \sum_{i=1}^{N_N^{ss}} 0.5(0.7 - \epsilon_i^{ss}) 300 k_B \left[ \left( \frac{r_{cry,i}}{r_i} \right)^{12} - 2 \left( \frac{r_{cry,i}}{r_i} \right)^6 \right], \end{aligned} \quad (S3)$$

where  $N_N^{bb}$ ,  $N_N^{bs}$  and  $N_N^{ss}$  denote the number of native contact pairs between backbone-backbone, backbone - side chain and side chain - side chain, respectively. The values of  $N_N^{bb}$ ,  $N_N^{bs}$  and  $N_N^{ss}$  are 172, 432 and 173, respectively.  $k_B$  is the Boltzmann constant,  $r_i$  denotes the distance

between  $i^{th}$  pair of beads, and  $r_{cry,i}$  denotes the corresponding distance in the crystal structure.  $\epsilon_h^{bb}$ ,  $\epsilon_h^{bs}$  and  $\epsilon_i^{ss}$  denote the strength of backbone - backbone, backbone - side chain and side chain - side chain interactions, respectively (Table S1). The values of  $\epsilon_i^{ss}$  are taken from the Betancourt-Thirumalai statistical potential [4].

The non-native interactions,  $E_{NB}^{NN}$ , are purely repulsive interactions and are given by

$$E_{NB}^{NN} = \sum_{i=1}^{N_{NN}} \epsilon_l \left( \frac{\sigma_i}{r_i} \right)^6 + \sum_{i=1}^{N_{ang}^{bb}} \epsilon_l \left( \frac{\sigma^{bb}}{r_i} \right)^6 + \sum_{i=1}^{N_{ang}^{bs}} \epsilon_l \left( \frac{\sigma_i^{bs}}{r_i} \right)^6 \quad (S4)$$

where  $N_{NN}(= 6973)$  is the total number of non-native interactions,  $N_{ang}^{bb}(= 62)$  is the number of pairs of backbone beads separated by 2 bonds in the SOP-SC model, and  $N_{ang}^{bs}(= 126)$  is the number of pairs of backbone and side chain beads separated by 2 bonds in the SOP-SC model.  $\sigma^{bb}$  is the diameter of the backbone beads, and  $\sigma_i^{bs}(= f[\sigma^{bb} + \sigma_i^{sc}]/2.0)$  is the sum of the radii of the backbone and the side chain in the  $i^{th}$  pair of angular interactions scaled by a factor  $f = 0.9$ . Values of the side chain radii are given in Table S2 in Ref.[5]

The values of the parameters used in the energy function (Table S1) are identical to the values previously used to successfully study the folding properties of the proteins GFP[5] and Ubiquitin[6]. We have used the same force-field to study the properties of different proteins, and as a result this force-field satisfies the criterion of a transferable force-field.

**Molecular Transfer Model:** To simulate Protein L folding thermodynamics and kinetics in the presence of Guanidine Hydrochloride we used the Molecular Transfer Model (MTM)[7, 8]. In the presence of a denaturant of concentration  $[C]$ , the effective coarse-grained force field for the protein using MTM is given by

$$E_{CG}(\{\mathbf{r}\}, [C]) = E_{CG}(\{\mathbf{r}\}, 0) + \Delta G_{tr}(\{\mathbf{r}\}, [C]), \quad (S5)$$

where  $E_{CG}(\{\mathbf{r}\}, 0)$  is given by Eq. S1,  $\Delta G_{tr}(\{\mathbf{r}\}, [C])$  is the protein-denaturant interaction energy in a solution with denaturant concentration  $[C]$ , and is given by

$$\Delta G_{tr}(\{\mathbf{r}\}, [C]) = \sum_{k=1}^N \delta g_{tr,k}([C]) \alpha_k(\{\mathbf{r}\}) / \alpha_{Gly-k-Gly}, \quad (S6)$$

where  $N(=N_{res} \times 2 = 128)$  is the number of beads in coarse-grained Protein L,  $\delta g_{tr,k}([C])$  is the transfer free energy of bead  $k$ ,  $\alpha_k(\{\mathbf{r}\})$  is the solvent accessible surface area (SASA) of the bead  $k$  in a protein conformation described by positions  $\{\mathbf{r}\}$ ,  $\alpha_{Gly-k-Gly}$  is the SASA of the bead  $k$  in

the tripeptide *Gly* – *k* – *Gly*. The radii for amino acid side chains to compute  $\alpha_k(\{\mathbf{r}\})$  are given in Table S2 in Ref.[6]. The experimental[7, 9, 10] transfer free energies  $\delta g_{tr,i}([C])$ , which depend on the chemical nature of the denaturant, for backbone and side chains are listed in Table S3 in Ref.[2]. The values for  $\alpha_{Gly-k-Gly}$  are listed in Table S4 in Ref.[2].

**Simulations:** The SOP-SC model of the polypeptide chain is simulated using Langevin dynamics at different temperatures ranging from 300 K to 430 K in low friction using the energy function given by eq. S1 to compute the average thermodynamic properties of the protein. The equations of motion are integrated using the equation

$$m\ddot{\vec{r}}_i = -\zeta\dot{\vec{r}}_i + \vec{F}_c + \vec{\Gamma}, \quad (S7)$$

where  $m$  is the mass of a protein beads,  $\zeta$  is the friction coefficient,  $\vec{r}_i$  is the position of the bead  $i$ ,  $\vec{F}_c = -\frac{\partial E_{TOT}}{\partial \vec{r}_i}$ ,  $\vec{\Gamma}$  is the random force with a white noise spectrum. The autocorrelation function of the random force in the discretised form is given by  $\langle \Gamma(t) \Gamma(t + nh) \rangle = \frac{2\zeta k_B T}{h} \delta_{0,n}$ , where  $n = 0, 1, \dots$  and  $\delta_{0,n}$  is the Kronecker delta function. The Langevin equation is integrated using the velocity Verlet algorithm[11, 12]. We used  $\zeta = 0.05 \text{ m}/\tau_L$  and  $h = 0.005 \tau_L$ , where  $\tau_L$  is the unit of time used to advance the simulation.

To compute thermodynamic properties of the protein in a denaturant solution of concentration  $[C]$ ,  $\Delta G_{tr}(\{\mathbf{r}\}, [C])$  is treated as perturbation to  $E_{CG}(\{\mathbf{r}\}, 0)$  in Eq. S5, and Weighted Histogram Method[7, 8, 13] is used to compute average value of various physical quantities at any  $[C]$ . The average value of a physical property  $A$ , at temperature  $T$ , and denaturant concentration  $[C]$  is computed using the equation

$$\langle A([C], T) \rangle = Z([C], T)^{-1} \sum_{k=1}^R \sum_{t=1}^{n_k} \frac{A_{k,t} e^{-(E_{k,t}(\{\mathbf{r}_{k,t}\}, [0]) + \Delta G_{tr}(\{\mathbf{r}_{k,t}\}, [C]))/k_B T}}{\sum_{m=1}^R n_m e^{f_m - E_{k,t}(\{\mathbf{r}_{k,t}\}, [0])/k_B T_m}}, \quad (S8)$$

where  $R$  is the number of simulation trajectories,  $n_k$  is the number of protein conformations from the  $k^{th}$  simulation,  $A_{k,t}$  is the value of the property of the  $t^{th}$  conformation from the  $k^{th}$  simulation,  $T_m$  and  $f_m$  are the temperature and free energy respectively from the  $m^{th}$  simulation,  $E_{k,t}(\{\mathbf{r}_{k,t}\}, [0])$  and  $\Delta G_{tr}(\{\mathbf{r}_{k,t}\}, [C])$  are the internal energy at  $[C] = 0$  and MTM energy respectively of the  $t^{th}$  conformation from the  $k^{th}$  simulation, and  $Z([C], T)$  is the partition function given by

$$Z([C], T) = \sum_{k=1}^R \sum_{t=1}^{n_k} \frac{e^{-(E_{k,t}(\{\mathbf{r}_{k,t}\}, [0]) + \Delta G_{tr}(\{\mathbf{r}_{k,t}\}, [C]))/k_B T}}{\sum_{m=1}^R n_m e^{f_m - E_{k,t}(\{\mathbf{r}_{k,t}\}, [0])/k_B T_m}}. \quad (\text{S9})$$

We performed Brownian dynamics simulations with the full Hamiltonian given by Eq. S5, and a friction coefficient, which approximately corresponds to that of water to study the burst-phase folding kinetics of Protein L. The equations of motion are integrated using the Ermak-McCammon algorithm[14],  $\vec{r}_i(t+h) = \vec{r}_i(t) + \frac{h}{\zeta} \vec{F}_c + \vec{\Gamma}$ . Here  $\vec{\Gamma}$  is a random displacement with a Gaussian distribution with mean zero and variance  $\langle \Gamma(h)^2 \rangle = \frac{2k_B T h}{\zeta}$ . The friction coefficient  $\zeta = 31.2 \text{ m}/\tau_H$  approximately corresponds to the value in water and, the value of  $h$  varies from  $0.001 \tau_H$  to  $0.01 \tau_H$  depending on the denaturant concentration. In the simulations, the characteristic unit of length  $a = 1 \text{ \AA}$ , energy  $\epsilon = 1 \text{ kcal/mole}$ , and mass  $m = 1.8 \times 10^{-22} \text{ g}$  (typical mass of the bead). The unit of time in Langevin dynamics simulations is  $\tau_L (= \sqrt{ma^2/\epsilon}) = 1.3 \text{ ps}$ . In Brownian dynamics, simulation time is mapped into real time,  $\tau_H$  using  $\tau_H \approx \frac{\zeta_H a^2}{k_B T} = \frac{(\zeta_H \tau_L / m) \epsilon}{k_B T} \tau_L \approx 47 \text{ ps}$ .

**Transition State Analysis:** 108 putative transition state structures (TSS) from the Langevin dynamics trajectory at  $T_M = 374.5 \text{ K}$  are identified using the conditions  $14 \text{ \AA} \leq R_g \leq 16.2 \text{ \AA}$  and  $0.5 \leq \chi \leq 0.6$  (Fig. S2) for the  $P_{fold}$  analysis. To compute  $P_{fold}$  for each putative TSS, 500 short simulation trajectories each of  $0.15 \mu\text{s}$  in length are initiated using the putative TSS as the initial conformation to compute the fraction of the trajectories, which land up in the NBA or the UBA (Fig. S6).



- 
- [1] C. B. Hyeon, R. I. Dima, and D. Thirumalai. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. Structure, 14:1633–1645, 2006.
- [2] Zhenxing Liu, Govardhan Reddy, Edward P O’Brien, and D Thirumalai. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. Proc. Natl. Acad. Sci. USA, 108(19):7787–7792, 2011.
- [3] J W O’Neill, D E Kim, D Baker, and K Y J Zhang. Structures of the B1 domain of protein L from *Peptostreptococcus magnus* with a tyrosine to tryptophan substitution. Acta Crystallogr. Sect. D-Biol. Crystallogr., 57(4):480–487, 2001.
- [4] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Prot. Sci., 8:361–369, 1999.
- [5] Govardhan Reddy, Zhenxing Liu, and D. Thirumalai. Denaturant-dependent folding of GFP. Proc. Natl. Acad. Sci. USA, 109:17832–17838, 2012.
- [6] Govardhan Reddy and D Thirumalai. Dissecting Ubiquitin folding using the Self-Organized Polymer model. J. Phys. Chem. B, 119(34):11358–11370, 2015.
- [7] E. P. O’Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. Proc. Natl. Acad. Sci. USA, 105:13403–13408, 2008.
- [8] Zhenxing Liu, Govardhan Reddy, and D. Thirumalai. Theory of the Molecular Transfer Model for Proteins with Applications to the Folding of the src-SH3 Domain. J. Phys. Chem. B, 116(23):6707–6716, 2012.
- [9] M. Auton and D. W. Bolen. Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. Biochemistry, 43:1329–1342, 2004.
- [10] E. P. O’Brien, B. R. Brooks, and D. Thirumalai. Molecular Origin of Constant  $m$ -Values, Denatured State Collapse, and Residue-Dependent Transition Midpoints in Globular Proteins. Biochemistry, 48:3743–3754, 2009.
- [11] T. Veitshans, D. Klimov, and D. Thirumalai. Protein folding kinetics: Timescales, pathways and

- energy landscapes in terms of sequence-dependent properties. Fold Des, 2(1):1–22, 1997.
- [12] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small clusters. J. Chem. Phys., 76:637–649, 1982.
- [13] S Kumar, J M Rosenberg, D Bouzida, R H Swendsen, and P A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. the method. J. Comput. Chem., 13:1011–1021, 1992.
- [14] D. L. Ermak and J. A. Mccammon. Brownian dynamics with hydrodynamic interactions. J Chem Phys, 69(4):1352–1360, 1978.

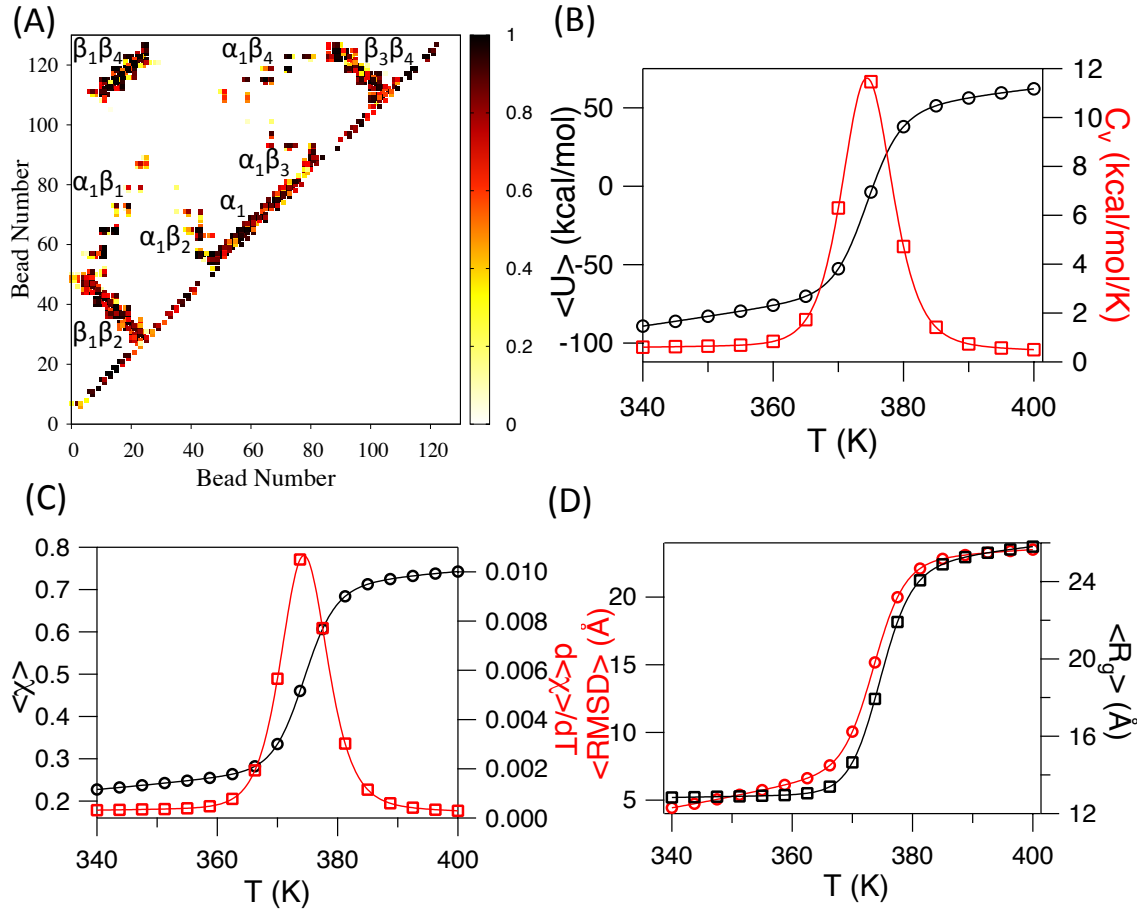


FIG. S1: (A) The contact map of Protein L shows contacts between various secondary structural elements in the folded state. (B) Average internal energy,  $\langle U \rangle$  (empty circles in black), and heat capacity,  $C_v$  (empty squares in red), as a function of temperature,  $T$ . (C) Structural overlap factor,  $\langle \chi \rangle$  (empty circles in black), and  $d\langle \chi \rangle/dT$  (empty squares in red), as a function of  $T$ . (D) Root mean square deviation,  $\langle RMSD \rangle$  (empty circles in red), and Radius of gyration,  $\langle R_g \rangle$  (empty squares in black), as a function of  $T$ .

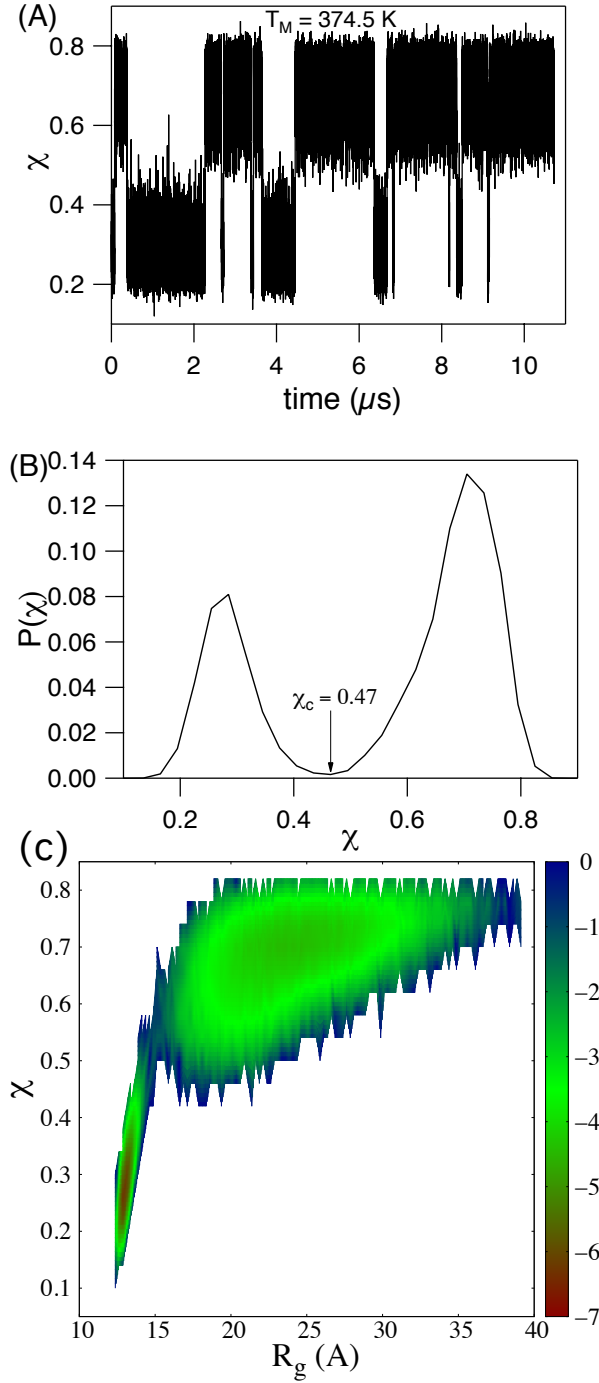


FIG. S2: (A) Structural overlap factor,  $\chi$ , plotted as a function of time at the melting temperature,  $T_M = 374.5 \text{ K}$ . (B) Probability distribution of  $\chi$ ,  $P(\chi)$ , at  $T_M$ . The value  $\chi_c = 0.47$  separates the unfolded basin of attraction (UBA) and native basin of attraction (NBA). (C) The free energy projected onto  $\chi$  and  $R_g$  using the relation,  $\Delta G = -k_B T_M \ln(P(R_g, \chi))$ , where  $P(R_g, \chi)$  is joint probability distribution of  $R_g$  and  $\chi$  at  $T_M$ , and  $k_B$  is the Boltzmann constant. The two basins corresponding to the UBA and NBA show two-state behaviour at  $T_M$ .

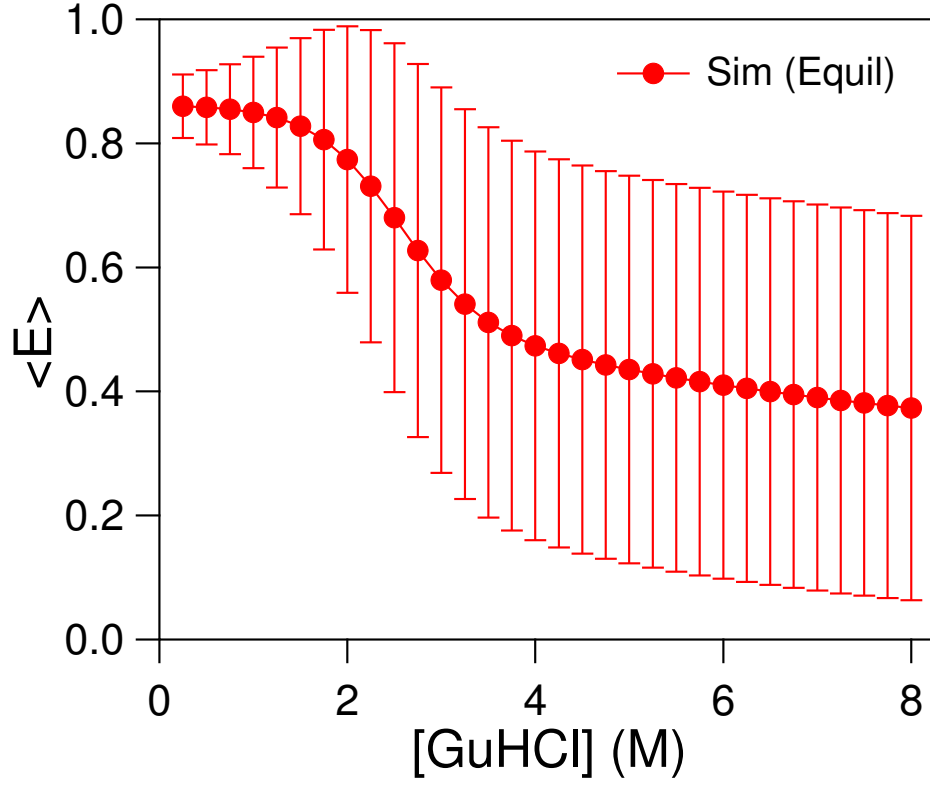


FIG. S3: Average FRET efficiency,  $\langle E \rangle$ , as a function of  $[GuHCl]$  at  $T = 357.7$  K.

TABLE S1: Parameters for the SOP-Side Chain model

Parameters	Protein
$R_o$	2.0 Å
$k$	20 kcal/mol/Å <sup>2</sup>
$R_c$	8 Å
$\epsilon_h^{bb}$	0.45 kcal/mol
$\epsilon_h^{bs}$	0.45 kcal/mol
$\epsilon_l$	1.0 kcal/mol
$\sigma^{bb}$	3.8 Å

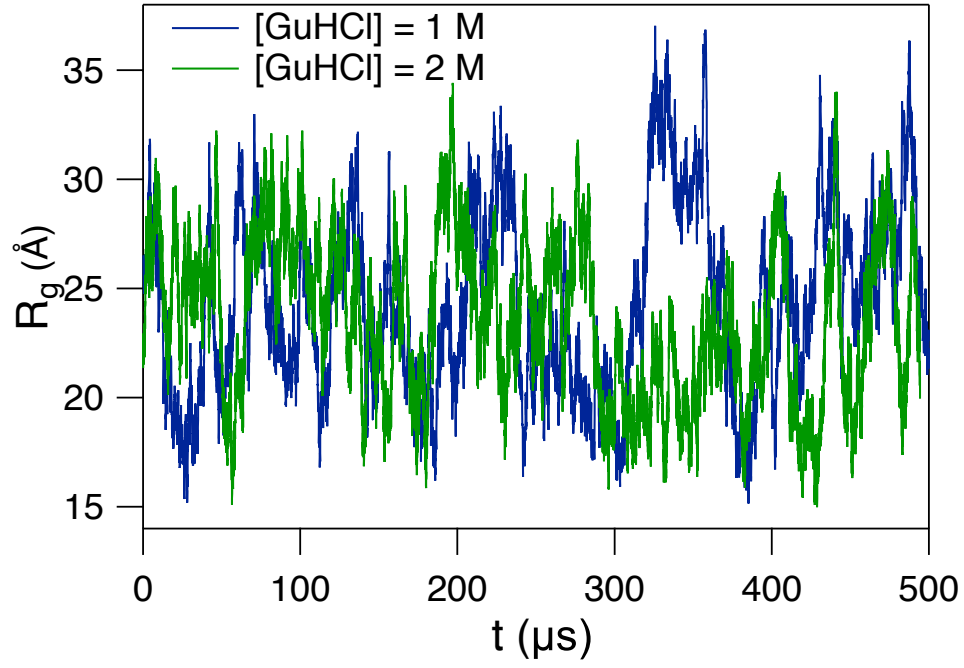


FIG. S4: Radius of gyration,  $R_g$ , plotted as a function of time,  $t$ , for Protein L folding trajectories in  $[GuHCl] = 1\text{ M}$  (blue) and  $2\text{ M}$  (green) at  $T = 357.7\text{ K}$ .

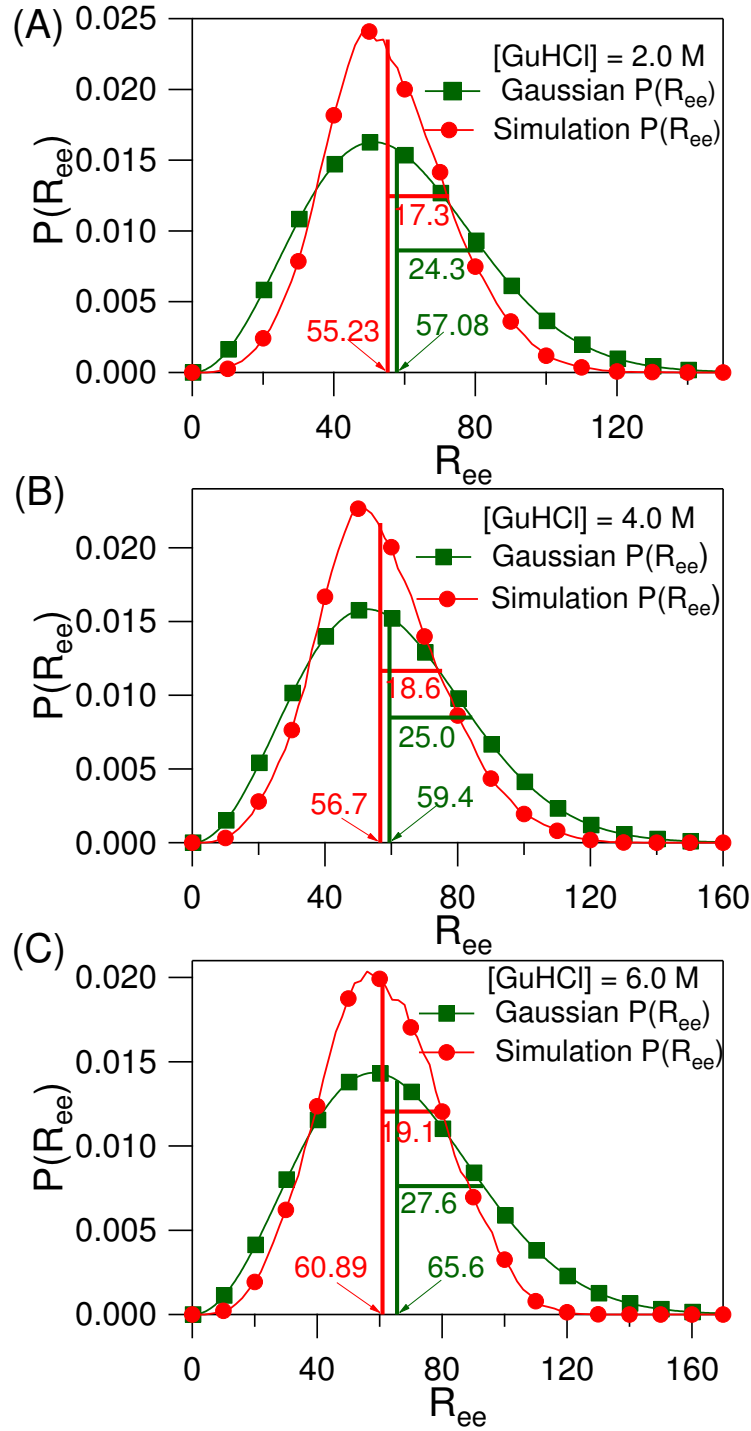


FIG. S5: The end-to-end distance,  $R_{ee}$ , probability distribution function  $P(R_{ee})$  during the burst phase (initial  $0.25\text{ ms}$ ) of protein L folding is in red circles.  $P(R_{ee})$  estimated from  $\langle E^{Burst} \rangle$  and Gaussian polymer chain statistics in green squares. (A)  $[GuHCl] = 2.0\text{ M}$ ,  $T = 357.7\text{ K}$  (B)  $[GuHCl] = 4.0\text{ M}$ ,  $T = 357.7\text{ K}$  and (C)  $[GuHCl] = 6.0\text{ M}$ ,  $T = 357.7\text{ K}$

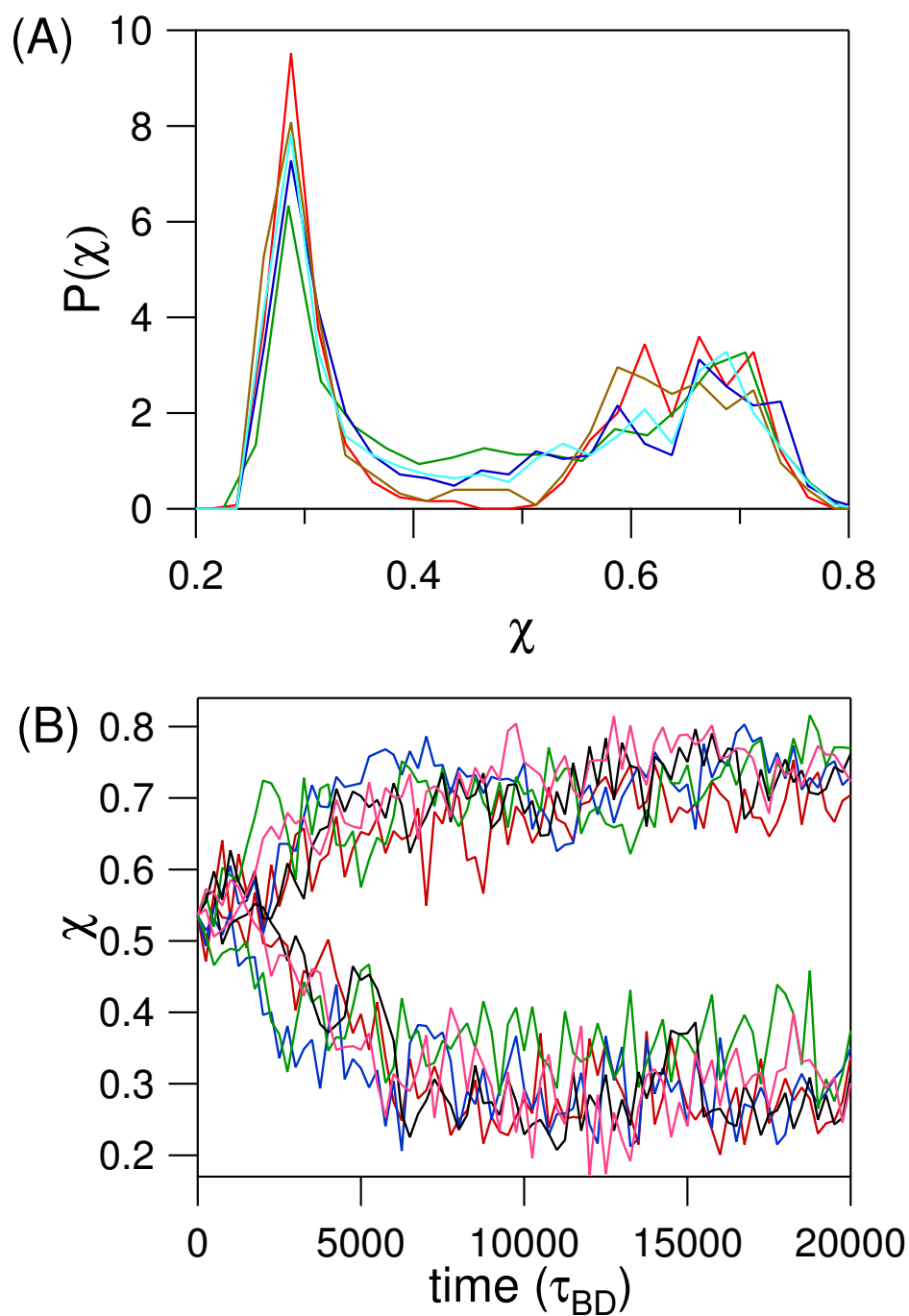


FIG. S6: (A) The distribution of the final structural overlap factor,  $\chi$ , for each transition state structure at the end of  $0.15 \mu s$  computed from 500 simulation trajectories. Data for 5 different structures is shown. (B) Simulation trajectories spawned using a transition state structure as the starting conformation land up in UBA and NBA.